UNIVERSITY OF CALIFORNIA, SAN DIEGO

Evolutionary Genetics of Self-incompatibility in Solanaceae and Papaveraceae

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Biology

by

Timothy Paape

Committee in charge:

Professor Joshua R. Kohn, Chair
Professor Ronald S. Burton
Professor Lin Chao
Professor Stephen G. Weller
Professor Christopher Wills

2009

The dissertation of Timothy Paape is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

Table of Contents

List of Tables

Chapter I

Chapter II

Chapter III

List of Figures

Chapter I

Chapter II

Chapter III

Acknowledgments

Thanks to Michael Lawrence, VE Franklin-Tong for paving the way for SI studies in poppies and to Gary Hannan for work with *Platystemon* and assistance with population locations. All of the UC Reserve staff that helped early on to find plant populations and assistance in seed collection, especially Mark Stromberg at Hastings and Daniel Dawson and Kim Rose at SNARL. To all of our collaborators (outside of the Kohn lab) on the bottleneck paper, Stacey Smith, Richard Olmstead, Lynn Bohs for sharing your data, samples and knowledge. Boris Igic and Matt Streisfeld were instrumental in orienting me in the lab as well as teaching me lots of great molecular skills. Special thanks to Sergei Kosakovsky Pond for sharing so much of your expertise, time and beer during our discussions of the comparative selection paper. Also Danny Wilson for responding to so many emails and sharing code for Omegamap and R. Thanks to Jill Miller for a great collecting trip down the Baja Pennisula to collect *Lycium*. Thanks to Ed Newbigin for all of the insight and into pollen S in Solanaceae.

I had the good fortune to be well funded throughout my research from many sources, they are: Jeanne Messier Memorial Fund, California Desert Research Fund, Mildred Mathias Fund, the National Science Foundation and to James Crow for his generous fellowship. Thanks to my advisor Josh Kohn for all of the advice (well most of it anyway), discussion, the lab space, grueling hours of editing and writing, letters of support, securing resources and encouragement over the past five years.

Thanks to my parents for all of the amazing support over the years. Thanks to my wife Amy for riding the roller coaster of events, both real and imaginary throughout my PhD. and for all of your comforting, care and love…and most of all, patience. A special thanks

to Yogananda for your garden and making Encinitas a wonderful sanctuary during my stay in San Diego. And of course all of the other blessings as well…

The text of Chapter I, in full, is a reprint of the material as it appeared in Molecular Biology and Evolution. I was the primary researcher and author. The co-authors listed in this publication contributed species level sequence data and plant material, and Dr. Kohn directed and supervised the research included in this chapter.

Curriculum vitae
Timothy Paape, Ph.D.

A) Professional Preparation:

2002 B.S. Biology- Fort Lewis College, Durango, CO

2003-2004 Lab Technician, Plant Sciences, University of Minnesota, St. Paul, MN

2009 Ph.D. Biological Sciences, University of California San Diego, CA

B) Appointments

2004-2009 Teaching Assistantship UC San Diego

C) Dissertation Subject

a) Genetics of S-locus polymorphism in three native California species (Papaveraceae)

b) Evolutionary genetic history and differential selection of S-RNases across the genera

D) Publications

Newbigin, E., **T. Paape,** J.R. Kohn. 2008. RNase based self-incompatibility: Puzzled by Pollen S. Plant Cell. 20:2286-2292

Paape, T., B. Igic, S. Smith, R. Olmstead, L. Bohs, J.R. Kohn. 2008. A 15-Million-Year-Old Genetic Bottleneck at the S-locus of the Solanaceae. Mol. Biol. Evol. 25: 655-663

Silverstein, K., M. Graham, **T. Paape**, and K.A. VandenBosch. 2005. Genome Organization of More Than 300 Defensin-Like Genes in Arabidopsis. Plant Physiology. 138: 600–610

Abstract of the Dissertation

Evolutionary Genetics of Gametophytic Self-Incompatibility in Solanaceae and Papaveraceae

by

Timothy Paape

Doctor of Philosophy in Biological Sciences

University of California, San Diego, 2009

Professor Joshua R. Kohn, Chair

Flowering plants are able to avoid inbreeding by several genetically based mechanisms. Gametophytic self-incompatibility (GSI) occurs when pollen is rejected in the style or on the stigma if it possesses a matching allele with either of the ovule parent's S-alleles. This mechanism typically involves a single genetic locus that is highly polymorphic within populations and species. S-alleles are maintained by strong negative frequency dependent selection that essentially favors alleles when they become rare in a population. This type of balancing selection preserves variation at the S-locus for millions of years enabling us to infer ancient demographic patterns through phylogenetic analyses of genealogies of S-alleles. GSI has been described in several taxa of Solanaceae but only one genus of Papaveraceae, the genus *Papaver*. Although the molecular

mechanisms of self-recognition in these respective families differ remarkably, the underlying theoretical predictions regarding their genetics and evolution are expected to be similar.

In **Chapter I** of this dissertation, I first explore the evolutionary history of a genetic bottleneck in the Solanaceae. Self-incompatible species in the sister genera *Physalis* and *Witheringia* share restricted variation at the S-locus indicative of an ancient bottleneck that occurred in a common ancestor. Using phylogenic approaches to look at S-allele variation in species of the subtribe Iochrominae, the clade containing *Physalis* and *Witheringia*, we are able to determine when this bottleneck event occurred. We then use two chloroplast markers, fossil calibrations and a Bayesian relaxed molecular clock approach to determine the approximate date of the bottleneck. In **Chapter II,** I examine the molecular evolution of individual codons from S-alleles from the bottlenecked lineages of *Physalis* compared to those of the non-bottlenecked lineages of *Solanum*. Because *Physalis* S-RNases appear to have diversified more recently than those of *Solanum*, we find significantly different patterns among amino acids undergoing positive selection using maximum likelihood phylogenetic and Bayesian coalescent methods. (increase in subst. at 4 fold degenerate sites or 3rd position (general synonymous relative to 1-2nd pos.) HyPhy.docs, overall dS increase in Physalis relative to Solanum; overall increase in dN in Physalis?)

In **Chapter III,** I explore the genetics of a putative S-locus polymorphism in three previously uncharacterized species of Papaveraceae native to California. Analyses of putative S-allele sequences from *A. munita* , *P. californicus* and *Romneya coulteri* sampled from natural populations have shown that each harbors substantial genetic

polymorphism homologous to stylar S-alleles from *Papaver rhoeas*. These genes appear to be expressed only in female reproductive tissues as expected for stylar S-locus products. In *A. munita* and *P. californicus*, greenhouse crosses among full sibs with matching putative S-genotypes usually don't result in seed set while crosses among individuals with non-matching genotypes almost always do. In addition, potential duplications at or near this locus have been detected in diploid *P. californicus*. Contrary to other well known SI systems, allelic genealogies from Papaveraceae show a general pattern of monophyletic clustering according to species. Genealogies of these species' S-alleles and those from newly sequenced *Papaver* alleles show general patterns of monophyletic clustering. The reduced levels of trans-specific polymorphism may be explained by founder events or population bottlenecks in each of the species, though other possibilities must also be considered. We also employ maximum likelihood models to estimate positive selection among putative alleles from these taxa.

Chapter I

A 15-Myr-Old Genetic Bottleneck

Abstract

Balancing selection preserves variation at the self-incompatibility locus (S-locus) of flowering plants for tens of millions of years, making it possible to detect demographic events that occurred prior to the origin of extant species. In contrast to other Solanaceae examined, self-incompatible species in the sister genera *Physalis* and *Witheringia* share restricted variation at the S-locus indicative of an ancient bottleneck that occurred in a common ancestor. We sequenced 14 S-alleles from the subtribe Iochrominae, a group that is sister to the clade containing *Physalis* and *Witheringia.* At least six ancient S-allele lineages are represented among these alleles demonstrating that the Iochrominae taxa do not share the restriction in S-locus diversity. Therefore the bottleneck occurred after the divergence of the Iochrominae from the lineage leading to the most recent common ancestor of *Physalis* and *Witheringia*. Using cpDNA sequences, three fossil dates, and a Bayesian relaxed molecular clock approach, the crown group of Solanaceae was estimated to be 51 MY old and the restriction of variation at the S-locus occurred 14.0 - 18.4 MY before present. These results confirm the great age of polymorphism at the S-locus and the utility of loci under balancing selection for deep historical inference.

Introduction

Theoretical population genetic studies of balancing selection predict that it will greatly increase the coalescence time of allelic polymorphism relative to neutral variation (Takahata 1990; Vekemans and Slatkin 1994). This prediction has been confirmed by studies of self-recognition loci such as the MHC loci of jawed vertebrates (Klein et al. 1993), and the mating compatibility loci of both fungi (Muirhead et al. 2002) and plants (Ioerger et al. 1990; Richman and Kohn 2000; Castric and Vekemans 2004). In all of these systems, the time to coalescence of allelic variation is far older than extant species. Loci under balancing selection can therefore provide evidence of historical genetic and demographic events that far predate current species, a utility that has been termed "molecular paleo-population biology" (Takahata and Clark 1993).

In many flowering plants, self-incompatibility (SI) systems allow hermaphroditic individuals to recognize and reject their own pollen in favor of pollen from other individuals, thus avoiding the deleterious effects of self-fertilization (de Nettancourt 1977). In single-locus gametophytic SI, as found in the Solanaceae (nightshade family) studied here, a match between the S-allele carried by the haploid pollen grain and either of the S-alleles in the diploid style triggers pollen tube rejection, preventing self-fertilization and also cross-fertilization if the cross-pollen grain carries either allele found in the female parent. In such systems, rare alleles have a selective advantage because they are compatible with more mates (Wright 1939). Selection favoring rare alleles is quite strong, even with large numbers of alleles segregating in populations. For instance, a new pollen S-allele entering a population that already contains 20 alleles has an 11.1% male mating advantage (Clark 1993).

Strong negative frequency-dependent selection is responsible for the two outstanding features of S-locus polymorphism. First, dozens of alleles occur in natural populations with alleles accumulating until a balance is reached between selection favoring rarity and drift causing allele loss (Wright 1939; Lawrence 2000). Second, alleles are often very old because, if any allele drifts towards rarity, selection acts to increase its frequency (Ioerger et al. 1990; Clark 1993). In the Solanaceae, the S-locus gene responsible for self-pollen recognition and rejection in the female tissue is an RNase (S-RNase hereafter; McClure et al. 1989). The great age of polymorphism at the S-locus is exemplified by the fact that S-RNase alleles from the same diploid individual of Solanaceae often differ at more than 50% of their amino acid sites. In addition, S-RNase alleles from species in different genera often cluster together in phylogenetic analyses, evidence of broadly shared ancestral polymorphism (Ioerger et al. 1990; Richman and Kohn 2000; Igic et al. 2004, 2006; Savage and Miller 2006). Much of the S-locus polymorphism found in SI Solanaceae was present in their common ancestor, which must also have been SI (Igic et al. 2004; 2006).

A striking contrast exists between the sequence diversity of S-alleles from species of the closely allied genera *Physalis* and *Witheringia*, and nearly all other Solanaceae whose S-alleles have been sampled (species of *Brugmansia*, *Lycium, Nicotiana, Petunia* and *Solanum*). While the numbers of S-alleles present in *Physalis* and *Witheringia* species are similar to those found in other Solanaceae (Lawrence

2000; Savage and Miller 2006; Stone and Pierce 2005; Igic et al. 2007), all 93 S-RNases sequenced from three *Physalis* (Richman et al. 1996a; Richman and Kohn 1999; Lu 2001) and two *Witheringia* (Richman and Kohn 2000; Stone and Pierce 2005) species cluster within only three S-allele lineages that pre-date the divergence of *Physalis* and *Witheringia*. For other Solanaceae, even small samples of alleles usually represent many more ancient lineages (reviewed in Richman and Kohn 2000; Castric and Vekemans 2004; see also Savage and Miller 2006; Igic et al. 2007). This finding has been interpreted as evidence of an ancient bottleneck that restricted variation at the S-locus in some common ancestor of the genera *Physalis* and *Witheringia.* No such restriction is evident at the S-locus of any other sampled SI Solanaceae (Richman et al. 1996b; Richman and Kohn 2000; Richman 2000; Igic et al. 2004; Stone and Pierce 2005; Igic et al. 2006) except for African species of the genus *Lycium* (Miller et al. in press), whose S-locus shows evidence of a bottleneck associated with colonization of the Old World from America. After the restriction of S-allele diversity in some common ancestor of *Physalis* and *Witheringia*, the remaining S-allele lineages diversified leaving the observed pattern of large numbers of S-alleles representing only a restricted number of ancient S-allele lineages.

In this paper we date the historical restriction of S-locus variation common to *Physalis* and *Witheringia*. First, we examine S-locus diversity in the South American monophyletic subtribe Iochrominae, which is found to comprise the sister group of the lineage containing *Physalis* and *Witheringia* by Olmstead et al. (in review). We ask whether the Iochrominae share the reduced set of S-allele lineages found in *Physalis* and *Witheringia*. If so, then the restriction of diversity at the S-locus predates the most recent common ancestor (MRCA) of the group containing the subtribe Iochrominae as well as *Physalis* and *Witheringia*. On the other hand, if the Iochrominae harbor a wide diversity of ancient S-allele lineages, then the restriction at the S-locus must have occurred after the divergence of the Iochrominae from the group containing *Physalis* and *Witheringia* but before the MRCA of *Physalis* and *Witheringia*. We then generated a cpDNA phylogeny of Solanaceae using a fossil-anchored Bayesian relaxed molecular clock approach, to date the branch along which the bottleneck is shown to have occurred.

Materials and Methods

Plant material and molecular procedures

Stylar tissue from one to four individuals from seven self-incompatible (Smith and Baum 2006a) species from the subtribe Iochrominae (*Dunalia brachyacantha* Miers, *Eriolarynx lorentzii* (Dammer) Hunz., *Iochroma australe* Griseb., *Iochroma cyaneum* (Lindl.) M. L. Green, *Iochroma gesnerioides* Miers, *Iochroma loxense* Miers, and *Vassobia breviflora* (Sendt.) Hunz.) was collected from plants growing at the University of Wisconsin greenhouse facility. Smith and Baum (2006b) determined these species SI status through manual self- and cross-pollinations. Seeds for these taxa were acquired largely from the Solanaceae Germplasm collection at Radboud University, Nijmegen, The Netherlands and a few from offspring of wild-collected individuals. Voucher numbers and accession information are given in Smith and Baum (2006a,b) and in the supplementary online material (http://mbe.oxfordjournals.org/). No large population samples were available for any single species within the Iochrominae. Sampling across species should provide an estimate of S-locus diversity within a group with the caveat that occasionally the same functional S-allele (same specificity) may be sampled from more than one species, making the estimate of the amount of S-allele diversity in the group conservative. Total RNA was extracted and reverse transcription performed to amplify S-alleles according to methods described by Richman et al. (1995), except for the application of 3'-RACE as in Igic et al. (2007). The forward degenerate primer PR1 (5'-GAATTCAYGGNYTNTGGCCNGA-3') amplifies from the 5' end of the conserved region C2 (Ioerger et al. 1991) to the 3'end of the coding region of the S-RNase cDNA. Products obtained via PCR were cloned using the TOPO TA Cloning Kit (Invitrogen Corp.) to separate alleles at the obligately heterozygous S-locus. Amplified cloned PCR products were screened by restriction digests (ten clones per individual, on average) and sent for automated sequencing by Eton Bioscience Inc., San Diego, CA.

Genealogy of S-alleles from Solanaceae

For phylogenetic analysis of S-RNase sequences from the Iochrominae, additional S-alleles were obtained from GenBank for the following species (number of alleles): *Lycium andersonii* (10), *Nicotiana alata* (6), *Petunia integrifolia* (6)*,* *Physalis cinerascens* (12), *Solanum carolinense* (9)*,* and *Witheringia solanacea* (15; see supplementary online material for Genbank accession numbers). We chose the taxa and allele sequences used for the phylogenetic analysis based on three criteria. First, we aimed for broad taxonomic representation across the Solanaceae. Second, S-RNase sequences had to cover at least the entire region between conserved regions two and five as described by Ioerger et al. (1991). Many sequences in GenBank are shorter and these were discarded. Third, in order to apply maximum-likelihood and Bayesian methods without prohibitively long computation times, we reduced the number of sequences used in the final dataset by first constructing a neighbor-joining tree in PAUP* v4.0b10 (Swofford 2002) using 71 S-allele sequences from Genbank along with our Iochrominae sequences. We then removed one of any intra-specific sister pair of non-Iochrominae alleles with fewer than ten amino acid differences. Twelve alleles were removed in this manner. This should not affect our goal of determining the number of ancient S-allele lineages represented among alleles recovered from the Iochrominae. Alleles that arose prior to the divergence of the Iochrominae are unlikely to fall between any very closely related pair of alleles from within a given species.

DNA sequences were manually aligned using BioEdit 7.0.1.4 (Hall 1999) and Se-Al 2.0 (Rambaut 1996) for phylogenetic analysis. Because many S-allele sequences in GenBank do not include the 3’ end of the gene, this region was removed from all sequences leaving 354 bp in the final alignment used for phylogenetic analysis. This represents approximately 62% of the coding region of the S-RNase gene including the hypervariable regions most frequently implicated as involved in specificity determination (Ioerger et al. 1991; Savage and Miller 2006; Igic et al. 2007).

We generated a maximum likelihood (ML) tree of S-alleles using PAUP* v4.0b10 (Swofford 2002). Maximum likelihood model parameters were determined using ModelTest 3.0 (Posada and Crandall 1998). The Akaike Information Criterion

(Akaike 1974) best fit model (TVM+I+ Γ) was used to heuristically search for the ML phylogeny. One S-RNase from *Antirrhinum hispanicum* (Plantaginaceae; Xue et al. 1996) was used as the outgroup. Bootstrap values were generated using a maximum-likelihood heuristic search of 1000 replicates using the same base frequencies found above to produce a 50% majority rule consensus tree.

We also used Mr. Bayes v3.1.1 (Ronquist and Huelsenbeck 2003) to generate a 50% majority consensus tree for comparison with the ML tree. Bayesian analysis was run using four simultaneous Markov chain Monte Carlo chains (3 heated and 1 cold) with a GTR+ Γ substitution model across sites. The analysis was run for 1,000,000 generations, sampling every 100th tree for a total of 10,000 trees. After determining stationarity, the initial 2501 trees were discarded from the burn-in phase. The remaining trees represent generations 250,001 to 1,000,000 (7500 trees) on which posterior probabilities were calculated.

Species phylogeny and divergence time estimation

The chloroplast sequences used for species divergence time estimation represent a subset of a much larger sample (200 species) of Solanaceae (Olmstead et al. in review). To reduce computational time needed for divergence time estimation, we limited our taxonomic sample to 29 Solanaceae representing only genera from which S-alleles have been sampled, or genera that represent basal nodes in the diversification of the Solanaceae ((e.g. *Schizanthus* and *Cestrum*, Olmstead and Sweere 1994, Olmstead et al. in review), but from which no S-locus information is currently available. An alignment of sequences from two chloroplast regions, *ndhF* coding (2116 bases) and *trnL-trnF* coding and intergenic spacer sequences (1377 bases) was used for a combined total of 3488 bases. For outgroup comparison and root placement we included *ndhF* and *trnL-trnF* sequence information from two species of Convolvulaceae (*Ipomoea batatas* and *Convolvulus arvensis*) which is considered to be the sister family to the Solanaceae (Olmstead and Sweere 1994). Because our small taxonomic sample could lead to erroneous estimation of relationships, we assumed the topological constraints (ordering of generic divergences) found in the larger phylogenetic analysis of Olmstead et al. (in review) which all receive ≥ 90% bootstrap support.

Likelihood ratio tests (Felsenstein 1988) were used to determine whether sequence data conformed to the expectation of a molecular clock. Maximum-likelihood models with and without the enforcement of a clock were performed using PAUP* on the constrained topology for each gene separately and on the combined dataset (both genes). The 2-parameter HKY85 model (Hasegawa et al. 1985) was selected with a four-category gamma distribution of rates across sites estimated from the data. Base frequencies, the transition/transversion ratio, and the gamma distribution shape parameter were estimated while running the ML analyses. The test statistic null model settings for each partition correspond to HKY85 + I + Γ + c with the alternative model being HKY85 + I + Γ assuming *N*-2 degrees of freedom where *N* is the number of terminal sequences. The distribution of likelihood ratio test under the hypothesis

$\Lambda = (-2[\ln_{\text{clock}} / \ln_{\text{without clock}}])$ was assumed to be as a χ^2.

Because the data do not conform to a strict molecular clock (see results), a Bayesian method (Thorne et al. 1998; Thorne and Kishino 2002; Drummond et al.

2006) of relaxing this assumption was used to estimate divergence times among species. The program BEAST v1.4 (Drummond and Rambaut 2003) performs both exponential and lognormal uncorrelated rate estimates of nucleotide substitution along lineages of a phylogeny using a Markov chain Monte Carlo simulation process. The Bayesian method of Drummond and Rambaut (2003) also allows the user to specify uncertainty in fossil dates using soft bound priors which is not possible using likelihood methods of divergence times (Yang 2006). Our prior probability parameters were as follows: we assumed the HKY85 + *Γ* model of nucleotide substitution with a proportion of invariant sites estimated from the sequence data. We fixed the mean substitution rate at the root node to be 0.0007 substitutions per million years, consistent with estimated coding and non-coding rates of cpDNA evolution (Palmer 1991; Schnabel and Wendel 1998). We assumed an uncorrelated lognormal relaxed model of rate heterogeneity among branches and a Yule prior model of speciation. The software also allows the user to calibrate specific nodes on the phylogeny to estimated fossil dates along with confidence intervals as priors. We used two fossil dates within the Solanaceae (*Solanum*-like and *Physalis*-like seeds from mid-Miocene and a Lower Eocene Convolvulaceae fossil, Benton 1993) as prior constraints of particular nodes (Table 1). Based on these fossils, we assumed normally distributed priors of 10 MY (SD = 4.0 MY) for the age of both *Solanum* and *Physalis* and a mean of 52 MY (SD = 5.2 MY) for the outgroup (Convolvulaceae) divergence (Magallón et al. 1999). The standard deviations on the priors represent the upper and lower bounds of the geological epochs from which the fossils were obtained.

We constrained the starting tree and all subsequent trees in the MCMC analysis to conform to the topology estimated by Olmstead et al. (in review). This preserves species relationships but allows for variation in node heights that translate to ages in millions of years. The MCMC was run twice each for 5,000,000 generations, sampling every 500th tree with a burn-in phase of 500,000 generations for each run. The two runs were checked for convergence and the posterior age distributions of the nodes of interest were analyzed using Tracer v1.3 (Rambaut and Drummond 2004). The estimated node ages for both runs were combined and re-sampled at a frequency of every 1000th tree, providing a sample of 10,000 trees. The time between the MRCA of *Physalis* and *Witheringia* and the MRCA of those genera plus the Iochrominae was estimated by subtracting the relevant node ages for each of the 10,000 samples. The results of the MCMC procedure are given as the mean and the 95% highest posterior density (HPD) intervals in millions of years. The mean and standard deviation of the duration of this branch were calculated from these values. Trees from both runs were combined to produce an ultrametric consensus tree using FigTree1.0 (Rambaut 2006). It should be noted that although the Bayesian program MULTIDIVTIME (http://statgen.ncsu.edu/thorne/multidivtime.html) does not allow soft-bound prior distributions on fossil dates, similar estimates for ingroup divergences were achieved using the above priors. We present the analysis using BEAST because it facilitates the estimation of the duration and associated error of the branch during which the restriction of variation at the S-locus occurred (see below).

Results

S-allele genealogy

A total of 14 different alleles from 15 individuals from the seven Iochrominae species were successfully amplified and sequenced. The low number of alleles relative to the number of individuals sampled resulted from two causes. First, several individuals shared common alleles. For example, four individuals of *Iochroma australis* and four of *Eriolarynx lorentzii* possessed only 3 different alleles per species. Our sample of plants was derived from small germplasm collections that likely contain lower S-locus diversity than would be found in nature. Second, only one allele was successfully isolated from two individuals.

As found in previous studies, the genealogy of Solanaceae S-alleles shows extensive shared ancestral polymorphism among most species (Figure 1B). The S-alleles of each species of *Petunia*, *Nicotiana*, *Lycium* and *Solanum* represent five to seven lineages that arose before the divergence of these genera. This is true even though only a subset of available alleles and species were included to simplify the analysis. In contrast, all alleles from *Physalis cinerascens* and *Witheringia solanacea* fall within only three lineages that pre-date the MRCA of these two genera. Previous studies that incorporated additional S-alleles and species have consistently found the same result (Richman et al. 1996; Richman and Kohn 2000; Lu 2001; Stone and Pierce 2005).

Despite the limited sampling of Iochrominae alleles, several observations can be made. First, in two cases, very similar alleles were recovered from different species of Iochrominae. These close pairs (E.lor1 and I.lox2, I.cya1 and I.ges2) differ by 2 and 3 amino acid residues, respectively, over the region compared and may represent sequence divergence within a specificity that arose after species divergence. Therefore, the 14 Iochrominae alleles sampled may represent fewer than 14 specificities. Among this set, we recovered at least six ancient Iocrhominae S-allele lineages five of which diverged from one another prior to the origin of the genus *Solanum* (Fig. 1). Alleles from group 1 (Fig. 1) are more closely related to alleles from *Solanum* than to other Iochrominae alleles. Given uncertainty in the topology in Fig. 1, this group of alleles could represent either one or two S-allele lineages that diverged prior to the origin of *Solanum*. Iochrominae S-allele lineages 2, 3 and 5 are each found to be sister to different S-alleles from *Nicotiana*, and group 6 is sister to a pair of alleles from *Petunia* and *Lycium*. Iochrominae S-alleles from group 4 are more closely related to alleles from *Physalis* and *Witheringia* than to alleles from other sampled genera. Only one Iochrominae S-allele (I.aus 2) falls within any of the three clades of alleles found in *Physalis* and *Witheringia*. The placement of that allele is uncertain; it may be sister to all other members of clade I (Fig. 1). A basal position for this allele would be consistent with diversification of this clade of alleles in *Physalis* and *Witheringia* after divergence of the Iochrominae. Most S-alleles recovered from Iochrominae fall neither within, nor sister to, the three clades of alleles represented in species of *Physalis* and *Witheringia*.

Species phylogeny and divergence estimates

Likelihood ratio tests strongly rejected the molecular clock for each chloroplast gene individually (χ^2 distributions: *ndhF*: Λ = 2[6867.69-6830.92]= 73.53, P < 0.001, 29 d.f.; *trnLF*: Λ = 2[4319.25-4211.59] = 215.32, P << 0.001, 28 d.f.) and for both genes combined (*ndhF* + *trnLF*: Λ = 2[11354-11247.94] = 212.84, P << 0.001, 31 d.f.). Using the program BEAST (Drummond and Rambaut 2003) we were able to relax the assumption of a strict molecular clock and determine an approximate time interval during which the bottleneck event occurred. The species phylogeny (Fig. 2) shows an estimated crown group age of 51 MY. The mean estimated age of the MRCA of the *Iochrominae*, *Physalis* and *Witheringia* was 18.4 MYA (95% HPD: 12.9, 24.2), while the mean estimated age of the MRCA of *Physalis* and *Witheringia* was 13.9 MYA (95% HPD: 9.6, 18.9). The mean difference between these two divergence times from 10,000 samples from the Bayesian analyses was 4.5 MY (SD = 2.10)

Discussion

Even the relatively small sample of S-alleles from the subtribe Iochrominae shows that they do not share the restriction in the diversity of S-allele lineages found in *Physalis* and *Witheringia*. Instead, they have a diverse set of S-alleles that comprise at least six lineages that pre-date the MRCA of the Iochrominae with *Solanum*. Only one S-allele sampled from the Iochrominae groups within any of the three S-allele clades found in *Physalis* and *Witheringia* and the position of this S-allele could be basal in that lineage. The diversity of S-allele lineages provides strong evidence that the restriction of S-locus variation common to *Physalis* and *Witheringia* occurred after divergence of the subtribe Iochrominae from the lineage leading to *Physalis* and *Witheringia*. Fossil-calibrated Bayesian relaxed molecular clock methods estimate the date of the restriction of diversity at the S-locus to between 14.0 and 18.4 MY before present.

Our analysis is in remarkable agreement with other evidence concerning the timing of S-locus restriction in *Physalis* and *Witheringia*. Richman (2000) used a lineage-through-time approach to show that diversification within the S-allele lineages found in *Physalis* began approximately one third of the way back from the present to the coalescence of all Solanaceae S-alleles. By comparison, S-alleles drawn from non-bottlenecked Solanaceae show a burst of diversification at the origin of S-allele genealogies, followed by a relative slowdown in diversification towards the present (Uyenoyama 1997; Richman and Kohn 2000; Savage and Miller 2006). Phylogenetic analyses have shown that the common ancestor of the Solanaceae possessed RNase-based SI (Igic and Kohn 2001; Steinbachs and Holsinger 2002; Igic et al. 2004, 2006). Our Bayesian estimate of the time of the MRCA of all extant Solanaceae is 51 MY ago (Figure 2) while the midpoint of the estimate for the time of the restriction of S-locus diversity is 16.2MYA, very close to one third (32%) of the way from the present to the MRCA of extant Solanaceae. Our dating results are also similar to those of Wikström et al. (2001) who used sequence data alone to estimate the age of the divergence of the Solanaceae from the Convolvulaceae as 65 (+/- 4) MY. However, in the absence of fossil data, their estimates of crown group and internal node ages for Solanaceae are somewhat younger than ours.

The timing of the loss of S-locus diversity in *Physalis* and *Witheringia* might be further narrowed down by examining S-allele diversity in the genus *Withania*. *Withania* and its close relatives have been placed sister to the group containing

Physalis and *Witheringia* in a recent molecular phylogenetic analysis using nuclear loci (Smith and Baum 2006a) but this group is placed sister to the clade containing the Iochrominae plus *Physalis* and *Witheringia* using a much larger taxonomic sample and the cpDNA regions used here (Olmstead et al. in review). Strong (> 90%) bootstrap support for alternative topologies in these studies may result from incongruity in the histories of nuclear and chloroplast loci, or from differences in taxon sampling. So far, no SI species of *Withania* have been reported (Kaul et al. 2005; Anderson et al. 2006). If one were found and shared the restricted number of S-lineages observed in *Physalis* and *Witheringia*, this would indicate both that *Withania* is more closely related to *Physalis* and *Witheringia* than are the Iochrominae and further constrain the window of time during which the restriction at the S-locus took place.

Restriction of sequence variation, but not S-allele number, in *Physalis* and *Witheringia* has been interpreted as resulting from a population bottleneck that severely reduced the number of S-alleles. Following the bottleneck, diversification of the remaining S-allele lineages restored S-allele numbers (Richman et al. 1996b; Richman and Kohn 2000; Richman 2000; Igic et al. 2004; Stone and Pierce 2005). Severe bottlenecks are required to reduce the number of lineages in a group to only three. Even a population of constant size 100 is expected to maintain six alleles at equilibrium (Wright 1939). Further loss of alleles after demographic recovery from a bottleneck is unlikely, due to the strong selection favoring maintenance of alleles when the population is below the equilibrium allele number. Maintenance of small population size over a protracted period decreases the time to coalescence, increasing the rate of loss of S-allele lineages. However, the size and duration of a population restriction needed to cause substantial turnover of S-allele lineages appears somewhat unrealistic. For instance, over the reasonable range of rates of origination of new S-alleles (10^{-6} – 10^{-9} per gene per generation; Vekemans and Slatkin 1994) the expected time to coalescence of S-allele variation in a population of constant size 100 ranges from 3×10^{5} to 1×10^{8} generations, respectively (Vekemans and Slatkin 1994). Thus either extremely long periods of reduced population size, or a brief but severe bottleneck, would be needed to cause substantial turnover of S-allele lineages.

A founder event in which extremely few, perhaps only two, individuals began a new population could have reduced S-allele number to the three lineages observed. However, if such an event were the cause, the new population would have had to maintain strict isolation from its source population and give rise to the genera *Physalis* and *Witheringia*. Any subsequent gene flow from the ancestral source population would almost certainly have increased the diversity of S-allele lineages above the three observed. In addition, any population founded with only three alleles would potentially suffer reproductive losses due to the fact that one in three potential mates would be incompatible.

Miller et al. (in press) provide the only example of a restriction in S-locus diversity in the Solanaceae outside of *Physalis* and *Witheringia*. In that case, the cause is almost certainly a founder event associated with long-distance dispersal. The genus *Lycium* (Solanaceae) is thought to have originated in South America (Levin and Miller 2005; Levin et al. 2007) but also occurs on several oceanic islands as well as southern Africa. African species form a monophyletic clade nested within the genus suggesting a single colonization event. A sample of S-alleles from African *Lycium* species contains significantly fewer lineages that predate the genus than similar samples of from New World species, suggesting a bottleneck associated with the colonization of Africa.

For *Physalis* and *Witheringia*, several non-demographic phenomena must also be considered that could have reduced the number of old S-allele lineages. First, a common ancestor of these two genera might have temporarily lost SI over most or all of its range, and then regained it after most S-allele diversity had been lost. Loss of SI is expected to lead to the collapse of S-locus polymorphism because it is no longer maintained by negative frequency-dependent selection (Igic et al. 2008). If all functional S-allele polymorphism is lost, the system cannot be regained because three alleles are needed for it to function or else all individuals would be mutually incompatible (Wright 1939). In addition, subsequent to the fixation of a mutation causing self-compatibility, loss of function mutations in other genes involved in the SI reaction typically arise (reviewed in Stone 2002; Igic et al. 2008), making the loss of SI essentially irreversible (Igic et al. 2006). However, Rick and Chetelat (1991) found that offspring of crosses among widely separated SC populations of the otherwise SI *Solanum habrochaites* (formerly known as *Lycopersicon hirsutum*) were restored to SI. Apparently, different mutations caused SC in the two populations. In situations such as this, it might be possible for a species to revert to SI if the selective forces acting on SI were to reverse before functional polymorphism at the S-locus was lost and before additional loss-of-function mutations accumulated. An additional possibility of this sort would be the restoration of SI in an SC species through interspecific hybridization. Such a scenario would have to involve extremely few hybridization events, however, because of the strong negative frequency-dependent selection favoring inter-specific transit of additional S-alleles.

Selective sweeps reduce variation and have been inferred in resistance loci such as the R-loci of plants (Bergelson et al. 2001) and the MHC of vertebrates (de Groot et al. 2002) thought to normally be subject to some form of balancing selection. For resistance loci, selection might at times be directional due to the prevalence of a certain disease for which one or a few alleles confer resistance. For the S-locus, such a scenario does not seem plausible. S-RNases are expressed only in stylar tissue and have no known function outside of incompatibility. Therefore conversion of balancing selection to directional selection favoring a particular allele seems unlikely.

Selective sweeps involving loci linked to the site of interest can also cause loss of variation. The S-locus comprises both pollen and stylar specificity-encoding loci, plus at least several other genes in a region of much reduced recombination (Stephan and Langley 1998; Wang et al. 2003; McClure 2006). However, it would be exceedingly difficult for directional selection on a linked locus to work against the strong force of negative frequency-dependent selection acting on the S-locus. Directional selection on a linked locus would have to be strong, and linkage extraordinarily tight for one or a few S-allele lineages to diversify and replace all others because of selection favoring a linked gene.

Finally, it is possible that certain S-allele lineages might diversify more rapidly than others leading to the loss of those that diversify more slowly. For instance, fewer amino acid substitutions might be needed to alter the specificity of alleles in some S-allele lineages. This explanation is unlikely given that the multiple remaining S-allele lineages in *Physalis* and *Witheringia* began diversifying at roughly the same time (Richman 2000; Igic et al. 2004). Uyenoyama (1997, 2003) suggested another factor that might affect the diversification rate of particular S-allele lineages. She noted that enforced heterozygosity at the gametophytic S-locus could shelter deleterious recessive mutations in genes linked to it. When a new allele arises, matings between it and its ancestral allele express this genetic load, lowering the fitness of both alleles until one or the other goes extinct leaving no evidence of diversification. Loss of load linked to certain S-alleles could lead to increased rates of diversification. Some experimental evidence for genetic load linked to the S-locus exists, at least for certain alleles (Stone 2004). However, Uyenoyama (1997) suggests that the loss of load linked to the S-alleles of an ancestor of *Physalis* and *Witheringia* may itself have been caused by fixation of deleterious recessives held in common by the few alleles remaining after a demographic bottleneck.

No scenario for the reduction in S-allele lineages seen in *Physalis* and *Witheringia* appears particularly persuasive or unequivocal. Whatever caused the loss of S-allele lineages in a common ancestor of *Physalis* and *Witheringia* represents an ancient and rare event during the estimated 50 MY diversification of the Solanaceae. Among taxa so far sampled, only the African clade of *Lycium* (Miller et al. in press) suffered a similar restriction. Because of the long duration of polymorphism at this locus, we can infer that no ancestor of any SI Solanaceae whose S-alleles do not show reduced numbers of ancient lineages suffered a historical restriction of S-allele diversity. These results provide strong evidence that events which occurred more than 10 MY ago can leave a persistent signature on loci under balancing selection. Inferring the precise cause of this restriction of S-locus diversity appears to be a considerably more difficult problem than documenting and estimating the time of occurrence.

Acknowledgements

Support for this work was provided by NSF DEB-0108173 and DEB-0639984 to J.R.K., DEB-0309184 to B.I. and J.R.K. and DEB-0316614 to L.B. and R.O., and NSF DEB-0309310 to S.D.S.

Table 1. Prior probability and posterior distribution estimates for calibration of the Solanaceae phylogeny. Priors are normal distributions based on fossils of seeds (Benton 1993) as follows: *Physalis*-like seeds from the mid-Miocene, *Solanum*-like seeds from mid to upper Miocene, *Convolvulus*-like seeds from the lower Eocene. Median values of particular epochs were used as mean dates (in millions of years = MY) for calibration points and each was issued an associated standard deviation and 95% confidence interval that was normally distributed based the upper and lower bounds of the epoch. The highest posterior density (HPD) values and their means from the 2 combined MCMC runs are reported in the third column. Posterior distribution values are the result of 2 runs of 5,000,000 generations each, sampled every 500th generation. These were combined into one log file and re-sampled at a frequency of every 1000th generation for a total of 10,000 trees.

Node Constrained	Normal Prior Distribution Mean, STDEV, (95% CI)	Posterior Distribution Mean, (95% HPD)
tMRCA *Physalis*	10 MY, 4.0, (3.4, 16.6)	11.9 MY, (7.9, 15.8)
tMRCA *Solanum*	10 MY, 4.0, (3.4, 16.6)	16.1 MY, (12.2, 20.6)
tMRCA Convolvulaceae and Solanaceae	52 MY, 5.2, (43, 60)	62.1 MY, (54.4, 69.7)

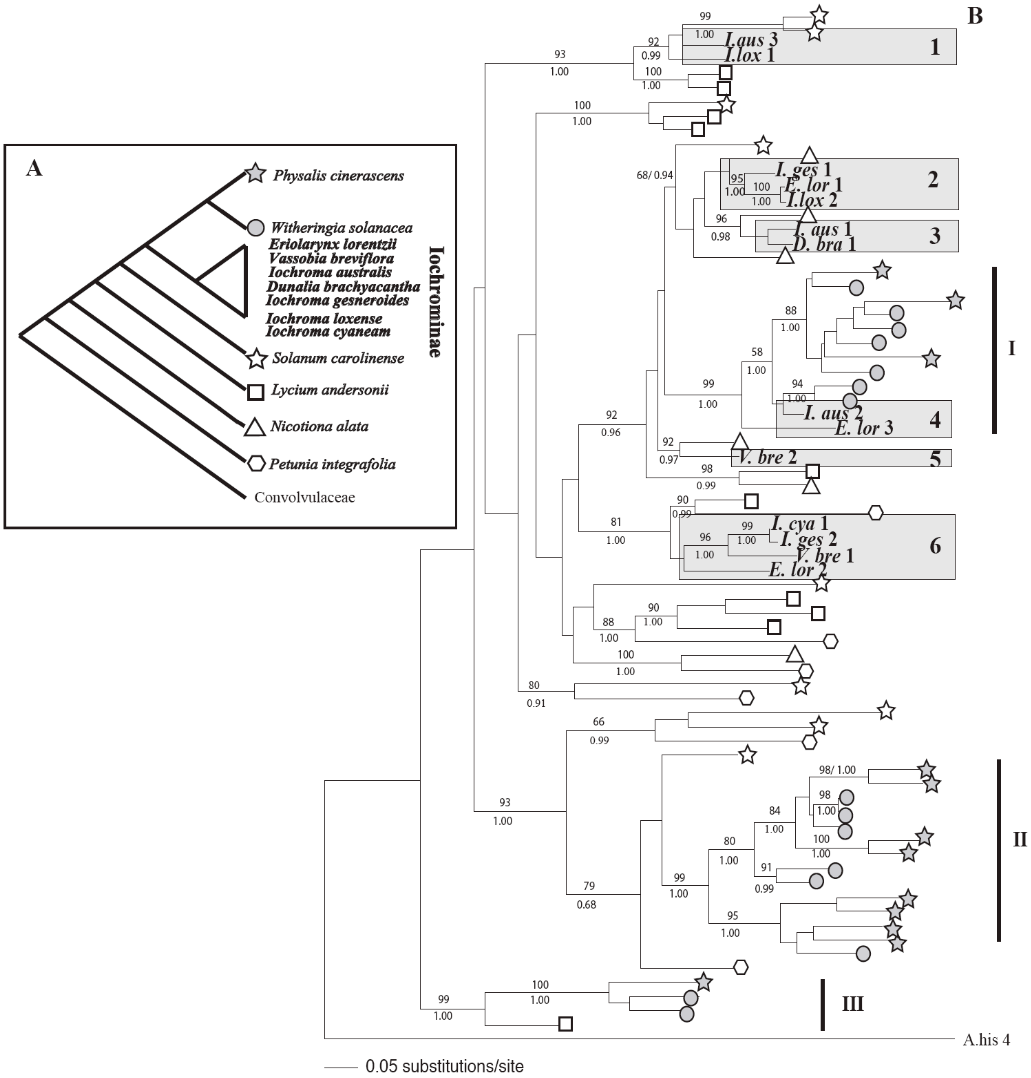

Figure 1. Maximum-likelihood phylogeny of 72 S-alleles from Solanaceae. Symbols correspond to alleles from taxa in the species phylogeny (inset A). The species phylogeny is redrawn from Olmstead et al. (in review). All nodes in the species phylogeny have >90% bootstrap support. All alleles from the genera *Physalis* and *Witheringia* are restricted to one of 3 lineages (indicated by Roman numerals) (B). The 14 alleles from the Iochrominae species (boldface) comprise at least six groups (indicated with arabic numerals and shaded boxes) that predate the Iochrominae. Five of these lineages (all but lineage 4) predate the divergence of *Solanum* from the other genera sampled. The S-allele phylogeny was constructed in PAUP* v4.0 (Swofford 2002). Bootstrap scores are indicated above branches and posterior probabilities > 80% generated by Mr. Bayes v3.0 (Ronquist and Huelsenbeck 2003) are below branches.

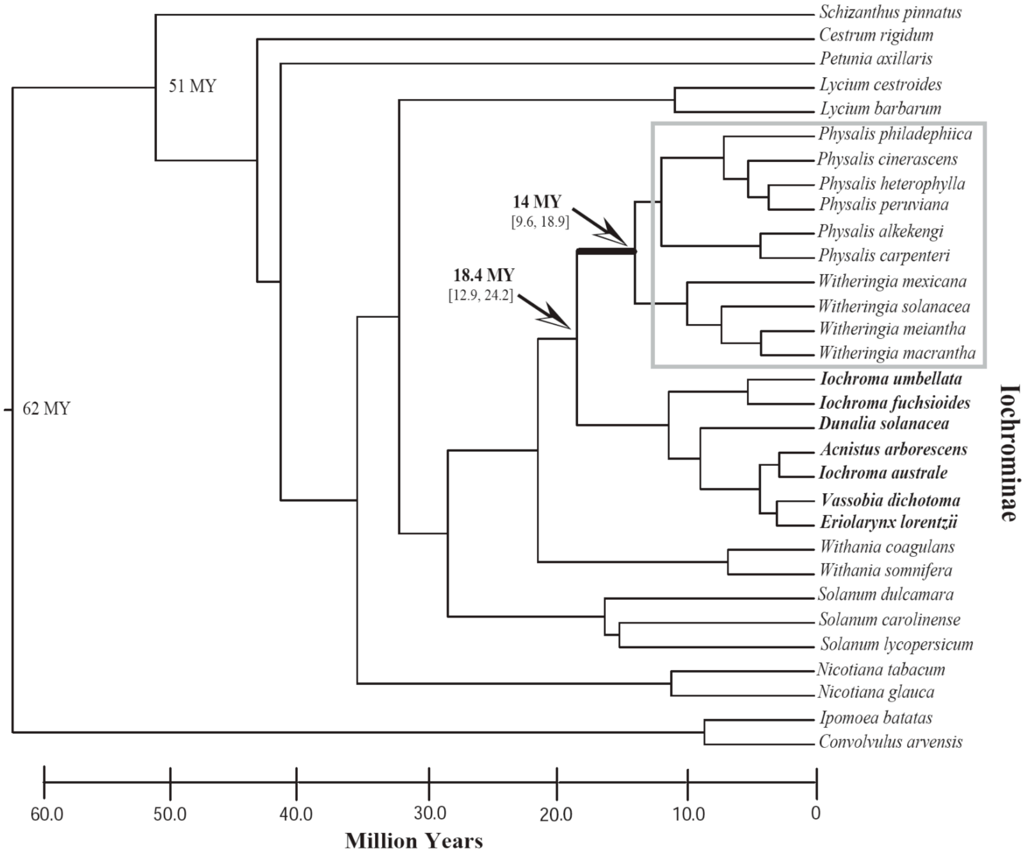

Figure 2. Bayesian consensus species phylogeny and divergence time estimates of the Solanaceae based on sequence data from two chloroplast genes. The root of the tree was estimated to be 62 million years (95% HPD: 54.4, 69.7 MY) and the crown group age was estimated at 51 million years (95% HPD: 38.6, 63.7) using BEAST v1.4 (Rambaut and Drummond 2003). HPD represents the 95% confidence intervals around the mean in millions of years. Reduction in the diversity of S-lineages in *Physalis* and *Witheringia* (See Figure 1) occurred between the two nodes indicated by arrows. The intervening branch is estimated to have a duration of 4.5 (SD = 2.10) MY.

References

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Aut. Contr. AC-19:716-716.

Anderson GJ, Bernardello G, Opel MR, Santos-Guera M, Anderson M. 2006. Reproductive biology of the dioecious Canary Islands endemic *Withania aristata* (Solanaceae). Am. J. Bot. 93:1295-1305.

Benton MJ. 1993. The Fossil Record 2. London: Chapman & Hall.

Bergelson J, Kreitman M, Stahl EA, Tian D. 2001. Evolutionary dynamics of plant R-genes. Science 292: 2281-2285

Castric V, Vekemans X. 2004. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Mol. Ecol. 13:2873–2889.

Clark AG. 1993. Evolutionary inferences from molecular characterization of self-incompatibility alleles. In: Takahata N, Clark AG, editors. Mechanisms of molecular evolution: introduction to molecular paleopopulation biology. Sunderland, MA: Sinauer Associates. p. 79-108.

de Groot NG, Otting N, Doxiadis GG, Balla-Jhagjhoorsingh SS, Heeney JL, van Rood JJ, Gagneux P, Bontrop RE. 2002. Evidence for an ancient selective sweep in the MHC class I gene repertoire of chimpanzees. Proc. Natl. Acad. Sci. USA 99:11748-11753

de Nettancourt D. 1977. Incompatibility in angiosperms. Berlin: Springer-Verlag.

Drummond AJ, Rambaut A. 2003. BEAST version 1.3. Available at http://evolve.zoo.ox.ac.uk/beast

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:699-710.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22: 521–565.

Hall TA 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98

Hasegawa M, Kishino Y, Yano Y. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160-174.

Igic B, Bohs L, Kohn JR. 2004. Historical inferences from the self-incompatibility locus. New Phytol. 161:97-105.

Igic B, Bohs L, Kohn JR. 2006. Ancient polymorphism reveals unidirectional breeding system shifts. Proc. Natl. Acad. Sci. USA 103:1359-1363.

Igic B, Lande R, Kohn JR. (in press) Loss of self-incompatibility and its evolutionary consequences. Int. J. Plant. Sci.

Igic, B., Smith WA, Robertson KA, Schaal BA, Kohn JR. (2007) Studies of self-incompatibility in wild tomatoes: I. S-allele diversity in *Solanum chilense* Dun. (Solanaceae). Heredity.

Ioerger TR, Clark AG, Kao T-h. 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc. Natl. Acad. Sci. USA 87:9732-9735.

Ioerger TR, Gohlke JR, Xu B, Kao T-h. 1991. Primary structural features of the self-incompatibility protein in Solanaceae. Sex. Plant Reprod. 4:81–87.

Kaul MK, Kumar A, Sharma A. 2005. Reproductive biology of *Withania somnifera* (L.) Dunal. Curr. Sci. 1375-1377.

Klein J, Satta Y, O'hUigin C, Takahata N. 1993. The molecular descent of the major histocompatibility complex. Annu. Rev. Immunol. 11:269–295.

Lawrence MJ. 2000. Population genetics of the homomorphic self-incompatibility polymorphisms in flowering plants. Ann. Bot. 85: 221–226.

Levin RA, Miller JS. 2005. Relationships within tribe Lycieae (Solanaceae): paraphyly of *Lycium* and multiple origins of gender dimorphism. Am. J. Bot. 92:2044–2053.

Levin RA, Shak JR, Miller JS, Bernardello G, Venter AM. 2007. Evolutionary relationships in tribe Lycieae (Solanaceae). Acta Hort. 745:225–239.

Lu, Y. 2001. Roles of lineage sorting and phylogenetic relationship in the genetic diversity at the self-incompatibility locus of Solanaceae. Heredity 86:195-205.

Magallón S, Crane PR, Herendeen PS. 1999. Phylogenetic pattern, diversity, and diversification of eudicots. Ann. Mo. Bot. Gard. 86:297–372

McClure BA. 2006. New views of S-RNase-based self-incompatibility. Curr. Opin. Plant Biol. 9:639-646.

McClure BA, Haring V, Ebert PR, Anderson MA, Simpson RJ, Sakiyama F, Clarke AE. 1989. Style self-incompatibility gene products of *Nicotiana alata* are ribonucleases. Nature 342: 955-957.

Miller JS, Levin RA, Feliciano NM. 2008. A tale of two continents: Baker's rule and the maintenance of self-incompatibility in *Lycium* (Solanaceae). Evolution (in press).

Muirhead CA, Glass NA, Slatkin M. 2002. Multilocus self-recognition systems in fungi as a cause of trans-species polymorphism. Genetics 161:633–641.

Olmstead RG, Bohs L, Abdel Magid H, Santiago-Valentin E, Collier SM, Garcia VF A molecular phylogeny of the Solanaceae. Taxonomy (in review).

Olmstead RG, Sweere JA. 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst. Biol. 43: 467–481.

Palmer JD. 1991. Plastid chromosome: structure and evolution. In: Bogorad L, Vasil IK, editors. The molecular biology of plastids. San Diego: Academic Press. p. 5–53.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Rambaut A. 1996. Se-Al: Sequence Alignment Editor. Available at http://evolve.zoo.ox.ac.uk/.

Rambaut A. 2006. FigTree. Available at http://evolve.zoo.ox.ac.uk/software/figtree.

Rambaut A, Drummond AJ. 2004 Tracer, version 1.1 [computer program] Available at http://evolve.zoo.ox.ac.uk/software.html?id=tracer

Richman AD. 2000. Evolution of balanced genetic polymorphism. Mol. Ecol. 9:1953-1963.

Richman AD, Kao T-h, Schaeffer SW, Uyenoyama MK. 1995. S-allele sequence diversity in natural populations of *Solanum carolinense* (Horsenettle). Heredity 75:405-415.

Richman AD, Uyenoyama MK, Kohn JR. 1996a. *S*-allele diversity in a natural population of ground cherry *Physalis crassifolia* (Solanaceae) assessed by RT-PCR. Heredity 76:497-505.

Richman AD, Uyenoyama MK, Kohn JR. 1996b. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science 273:1212–1216.

Richman AD, Kohn JR. 1999. Self-incompatibility alleles from *Physalis*: implications for historical inference from balanced genetic polymorphisms. Proc. Natl. Acad. Sci. USA 96:168–172.

Richman AD, Kohn JR. 2000. Evolutionary genetics of self-incompatibility in the Solanaceae. Plant Mol. Biol. 42:169-179.

Rick CM, Chetelat R. 1991. The breakdown of self-incompatibility in *Lycopersicon hirsutum*. In: Hawkes L, Nee M, Estrada N, editors. Solanaceae III: Taxonomy, chemistry, evolution. Richmond, UK: Royal Botanic Gardens Kew and Linnean Society of London. p. 253-256.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Savage, AE, Miller JS. 2006. Gametophytic self-incompatibility in *Lycium parishii* (Solanaceae): allelic diversity, genealogical structure, and patterns of molecular evolution. Heredity 96:434-444

Schnabel A, Wendell JF. 1998. Cladistic biogeography of *Gleditsia* (Leguminosae) based on *ndhF* and *rpl16* chloroplast gene sequences. Amer. J. Bot. 85:1753-1765.

Smith, SD and Baum DA. 2006a. Floral diversification and pollination biology for the Andean clade Iochrominae (Solanaceae). Am. J. Bot. 98: 1140-1153

Smith, SD, Baum DA. 2006b Systematics of Iochrominae (Solanaceae): Patterns in floral diversity and interspecific crossability. Acta Horticulturae 745: VI International Solanaceae Conference : Genomics Meets Biodiversity

Stephan W, Langley CH. 1998. DNA Polymorphism in *Lycopersicon* and crossing-over per physical length. Genetics 150:1585-1593.

Steinbachs JE, Holsinger KE. 2002. S-RNase-mediated gametophytic self-incompatibility is ancestral in eudicots. Mol. Biol. Evol. 19:825-829.

Stone JL. 2004. Sheltered load associated with S-alleles in *Solanum carolinense*. Heredity 92: 335–342.

Stone JL, Pierce SE. 2005. Rapid recent radiation of S-RNase lineages in *Witheringia solanacea* (Solanaceae). Heredity 94:547-555.

Swofford DL. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, MA:Sinauer Associates, Sunderland, Massachusetts.

Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc. Natl. Acad. Sci. USA 87:2419–2423.

Uyenoyama MK. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics 137: 1389–1400.

Uyenoyama MK. 2003. Genealogy dependent viability among self-incompatibility genotypes. Ther. Pop. Biol. 63: 281–293.

Vekemans X, Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics 137:

Takahata N, Clark AG. 1993. Mechanisms of molecular evolution: introduction to molecular paleopopulation biology. Sunderland, MA: Sinauer Associates.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Bio. Evol. 15:1647-1657.

Thorne JL, Kishino H. 2002. Divergence time estimation and rate evolution with multilocus data sets. Syst. Biol. 51:689-702.

Uyenoyama MK. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics 147:1389-1400.

Vekemans X, Slatkin M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics 137:1157–1165.

Wang Y, Wang X, McCubbin AG, Kao T-h. 2003. Genetic mapping and molecular characterization of the self-incompatibility (S) locus in *Petunia inflata*. Plant Mol. Biol. 53:565-580.

Wikström, N, Savolainen, V, Chase, MW. 2001. Evolution of the angiosperms: calibrating the family tree. Proc. R. Soc. B. 268: 2211-2220.

Wright, S. 1939. The distribution of self-sterility alleles in populations. Genetics 24:538-552.

Xue Y, Carpenter R, Dickinson HG, Coen ES. 1996. Origin of allelic diversity in *Antirrhinum* S locus RNases. Plant Cell 8:805-814.

Yang, Z. 2006. Molecular Clock and Species Divergence Times in Computational Molecular Evolution. Pp 248-258. Oxford University Press.

The text of Chapter I, in full, is a reprint of the material as it appeared in Molecular Biology and Evolution. Paape, T., B. Igic, S. Smith, R. Olmstead, L. Bohs, J.R. Kohn. 2008. A 15-Million-Year-Old Genetic Bottleneck at the S-locus of the Solanaceae. Mol. Biol. Evol. 25: 655-663. I was the primary researcher and author. The co-authors listed in this publication contributed species level sequence data and plant material, and Dr. Kohn directed and supervised the research included in this chapter.

Chapter II

Strength of positive selection of S-RNases in *Physalis* and *Solanum* (Solanaceae)

Abstract

The S-RNase system of Solanaceae is a highly polymorphic molecular self-incompatibility gene that has different evolutionary histories among genera. Strong balancing selection has revealed genealogical characteristics that are unique in the genus *Physalis* where shared ancestral polymorphism is reduced to three lineages due to a historical bottleneck. In sharp constrast, the genus *Solanum* did not undergo a historical reduction in population size evident by many trans-generic lineages within Solanaceae. We used maximum likelihood phylogenetic and Bayesian coalescent methods to detect differential positive selection using ratios of non-synonymous (dN) and synonymous substitution (dS) rates at individual codons. Comparisons of 47 *Physalis* and 48 *Solanum* S-RNase distributions of dN and dS rates show that the former has experienced significantly greater strengths of positive selection and a significantly greater selective regime. Highest posterior densities of dN/dS (ω) at individual codons were used to determine the ratio of ω values using a Bayesian coalescent method to estimate selection. Values of ω significantly greater than unity suggest that dN/dS is greater at given sites for *Physalis* for 36 amino acids. A nested maximum likelihood method that estimates differential selection at codons appeared considerably less powerful at detecting selection than the coalescent method with only 17 sites still favoring greater selection in *Physalis.* Both methods predicted the greatest differences in strengths of selection in the most variable regions of the S-RNases, as expected if these regions are responsible for functional changes in self-recognition. While average dN is nearly double for *Physalis* relative to *Solanum* S-alleles but average dS is an order of magnitude lower in the former. This is due to the

greater age of Solanum S-allele diversity and this accumulation of synonymous substitutions effects comparative estimates of dN/dS ratios.

Introduction

Highly polymorphic genes undergoing diversification can leave distinct signatures of positive selection across populations or taxa. Self-incompatibility (SI) polymorphisms are maintained by balancing selection over long evolutionary time scales. The selective advantage of an S-allele is inversely proportional to its frequency in the population (Wright 1939; Clark 1993). Rare alleles tend to increase in frequency because they are compatible with more potential mates. Diversifying selection leads to genealogical relationships among S-alleles that reflect their long persistence. Shared ancestral polymorphism is commonly observed where alleles from different species and genera cluster together in phylogenetic reconstructions (Richman et al. 1996; Kusaba et al. 1997; Sonneveld et al. 2001). This implies that S-alleles are often much older than the species from which they are sampled. Coalescence times of S-locus polymorphisms are often estimated as a few tens of millions of years, far longer than coalescence times of polymorphism at standard loci (Beschgaard et al. 2006; Paape et al. 2008).

Richman et al. (1996) detected a remarkable reduction in the extent of shared-ancestral polymorphism among S-alleles sampled from *Physalis crassifolia* relative to most other Solanaceae. In particular, *P. crassifolia* alleles, while numerous, all belonged to just three trans-generic lineages while alleles sampled from most other Solanaceae represented far more ancient lineages. For instance, estimates of historical effective population sizes of *Solanum carolinense* and *P. crassifolia* based on the number of extant alleles (*n* = 13 and *n* = 28 respectively) and ancient lineages (*k* = 9 and *k* = 3 respectively) showed at least an order of magnitude decrease in *Physalis* relative to *Solanum* (Richman et al. 1996). The genealogical pattern of alleles representing only three ancient lineages in *Physalis crassifolia* was found to be shared by other SI *Physalis* species and by and SI members of the closely related genus *Witheringia* (Richman and Kohn 1999; Lu 2001, 2005; Igic et al. 2004, 2007). These findings have been interpreted to be the result of a historical S-locus bottleneck in a common ancestor of *Physalis* and *Witheringia* that is not shared with *Solanum* or other major genera of Solanceae (Richman et al. 1996; Richman 2000; Paape et al. 2008).

Genealogical patterns revealed by phylogenies all suggest that the restricted lineages containing *Physalis* S-alleles underwent a rapid diversification process following the bottleneck event (Richman and Kohn 1999; Richman 2000; Stone and Pierce 2005) beginning approximately 15 million years ago (Paape et al. 2008). This provides an opportunity to examine patterns of diversifying selection on S-locus coding sites between taxa that have unique evolutionary histories. The more recently diversified S-alleles of *Physalis* relative to other taxa, might be expected to show greater rates of non-synonymous substitutions because of the strength of recent diversifying selection. In addition, the sites under positive selection thought to be those that confer specificity, might differ between the lineages found in *Physalis* and those from other taxa as it is not clear whether the same sites are under selection in all S-lineages.

Here we restrict our study to the stylar specificity component of the self-incompatiblity system of Solanaceae. The stylar gene product is an extracellular RNase (S-RNase) that is necessary and sufficient for self-pollen recognition and rejection. Positive selection has been estimated among S-RNase codons (Takebayashi et al. 2003; Savage and Miller 2006; Nunes et al. 2006; Igic et al. 2007) using various methods, most commonly maximum likelihood phylogenetic approaches proposed by Nielsen and Yang (1998) and more recently by coalescent based methods described by Wilson and McVean (2006; Vieira et al. 2007). Recent variations of phylogenetically based methods that estimate sites under selection (Kosakovsky Pond and Frost 2005) have also been applied to datasets encoding homologous viral genes that have undergone differential selection at the population level (Kosakovsky Pond et al. 2006).

Exceptionally high within species S-allele sequence polymorphism allows us to investigate positive selection on amino acids among S-RNases both within and across genera. *Physalis* and *Solanum* (Solanaceae) S-loucs polymorphisms provide potentially useful contrasts because diversification at the S-locus in the different genera took place during different time periods and among different S-allele lineages. The post bottleneck diversification of S-alleles in *Physalis* presents an opportunity to examine contrasting signatures of positive selection at the codon level. Although Savage and Miller (2006),Vieira et al. (2007), and Miller et al. (2008) have identified different positively selected sites among S-RNases, no studies of S-alleles test whether two datasets have statistically different selective pressures along the gene or at individual codons. Lineage selection models ('branch-site' models) described by

Zhang et al. (2005) are often biologically unrealistic because they assume a null model of uniform selection at all other branches in a phylogeny outside of the branch or lineage of interest (re-address this in Discussion for future development). Rather than isolating individual branches from within a phylogeny containing alleles from both genera, we treated alignments and phylogenies of each as distinct datasets. We use a series of nested and Bayesian models to answer the questions: 1) are the distributions of dN/dS, proportions of dN/dS at sites, and the strength of selection significantly different between genera? 2) Do the sites under selection differ among the genera and do selected sites occur in hyper-variable regions and/or conserved regions? 3) Are these differences due to higher dN or dS in one dataset relative to the other? 4) Has there been a significant increase in average substitution rates (ie. mutation) over time between representative species of each genus?

While countless studies have identified coding sites that appear to be under selective constraint or positive selection, few studies to date have employed methods to determine whether sequences from different populations or species show statistically different patterns of selection on codons (but see Kosakovsky Pond et al. 2006 and Blais et al. 2008). We present some current methodology used to analyze different strengths of divergent selection among S-alleles and discuss their application toward other potential systems as well as the need for new methods or modifications of existing ones.

Materials and Methods

Sequences and Phylogeny Construction

Amino acid and nucleotide S-RNase sequences were obtained from GenBank for 12 *Physalis cinerascens*, 36 *Physalis longifolia*, 17 *Solanum carolinense*, 32 *Solanum chilense* alleles and one *Antirrhinum hispanicum* (Ahis5) allele as an outgroup sequence. Automated alignment of the complete dataset containing all S-alleles was performed using ClustalX (Thompson et al. 1997) and manually aligned using Se-Al v2.0 (Rambaut 2002). A nucleotide alignment was matched with corresponding amino acids to produce a codon alignment using PAL2NAL (Suyama et al. 2006) that resulted in 131 codons. A phylogeny of all S-alleles (n = 98) was created using Mr. Bayes v3.1 (Ronquist and Huelsenbeck 2003) to generate a 50% majority consensus topology. The analysis was run under a GTR+ Γ + I substitution model for 1,000,000 generations, sampling every 100th tree for a total of 10,000 trees. The initial 2501 trees were discarded from the burn-in phase. The remaining trees represent generations on which posterior probabilities were calculated.

Separate datasets were compiled for each genus: one that contained 48 *Physalis* (both *P. cinerascens* and *P. longifolia*) and the other with 49 *Solanum* (both *S. carolinense* and *S. chilense*) S-alleles. Corresponding topologies for each dataset were pruned from the Bayesian consensus tree using TreeEdit v1.0a10 (Rambaut and Charleston 2001) to maintain genealogical relationships that are reflected when all taxa's alleles are included. Although each dataset represents samples of alleles from 2 species of each genus, the genealogical relationships are expected to be maintained within genera because of trans-specific polymorphism. The same tree topology for each dataset was used in all subsequent selection analyses that utilize phylogenies. A general time reversible (GTR) model of nucleotide substitution is used for all subsequent phylogenetic selection analyses so that direct comparisons can be made across models and datasets. Pairwise nucleotide divergence π was estimated for synonymous and non-synonymous substitutions for all taxa using DNASP 4.0 (Rozas et al. 2003). Sequence alignments, Newick string tree topologies and HYPHY likelihood functions for the *Physalis* and *Solanum* datasets can be found as Nexus files in online Supplementary data XXXX (Nexus files).

Selection Estimates

Distribution of dN and dS Rates

The most general test of the relative strength of selection across two datasets compares the distribution of synonymous and non-synonymous substitution rates using a random effects likelihood (REL) approach (Kosakovsky Pond et al. 2009) implemented in the program HYPHY (Kosakovsky et al. 2005). This consists of several nested models for hypothesis testing that begins by creating 4 bin general discrete distributions of rate classes for each dataset similar to the likelihood ratio tests (LRTs) described by Nielsen and Yang (1998) and implemented in PAML (Yang 2000). Rate classes are as follows: two bins for negative selection where $dS_1 > dN_1$ and $dS_2 > dN_2$; one bin for neutral evolution $dS_3 = dN_3$; and one for positive selection $dS_4 < dN_4$. Null hypotheses comparing both datasets are as follows: a) H_0: $dS_1/dN_1= dS_2/dN_2$ for the same strength of selection, b) H_0: $p_1 = p_2$ for the same proportion of positively selected sites, c) the same selective regime which combines both a) and b) (H_0: $dS_1/dN_1 = dS_2/dN_2$ and $p_1 = p_2$), and finally d) H_0: distribution of rates in both datasets are the same. An independent distribution model of rates that are free to vary for both datasets is set as the alternative hypothesis to which the null model likelihoods (a, b, c and d) are tested. Models are rejected by -2ΔlnL (ΔlnL = the difference in log likelihoods of the two models) where significance is determined by χ^2 distribution with the degrees of freedom (df) equal to the difference in the number of parameters between models.

Paml 3.15

To estimate the ratio (ω) of non-synonymous (d_N) to synonymous (d_S) substitutions at various amino acids we first used the program *codeml* in PAML 3.15 (Yang 2000). Values of $\omega < 1$ for individual codons indicates purifying selection while sites with $\omega = 1$ are considered neutral. Positive selection at the amino acid level is predicted when $\omega > 1$. A series of nested neutral and selection models first developed by Nielsen and Yang (1998) and Yang et al (2000) use likelihood ratio tests (LRT) to determine the model that best fits the data. The null model M1 (neutral) constrains all sites to be either of class $\omega = 0$ or $\omega = 1$ while the alternative model M2 (selection) adds a third class in which ω is free to vary at individual sites. Model M3 (selection) assumes three discrete site classes ($\omega_0 < 1$, $\omega_1 = 1$ and $\omega_2 >1$) with three corresponding proportions (p_0, p_1, p_2) estimated from the data. Models are then compared and rejected as described in the Distribution of Rates section above. Sites estimated to be under positive selection are determined by an empirical Bayes approach (Yang et al. 2005) where posterior probabilities are estimated from rates within each site class. Because we are primarily concerned with comparing posterior probabilities from the robust general discrete (M3) model with a subsequent coalescent analysis, we forgo full analyses including models with more complex rate distributions (ie. M7 and M8).

OmegaMap v0.5

Wilson and McVean (2006) developed a coalescent method that is able to simultaneously estimate dN/dS (ω) at individual codons and the population recombination rate along DNA sequences. The utility of this method was first investigated when samples of sequences have been found to undergo recombination, thus yielding more than one possible phylogenetic tree. They implemented a population genetics likelihood approximation (described in McVean et al. 2002) to the coalescent to infer recombination and extended it to estimate ω in the program OmegaMap v0.5. The model of base substitution including transition/transversion rates among codons was adopted from Nielsen and Yang (1998). Rather than using a maximum likelihood approach to estimate the selection parameter, OmegaMap employs a Bayesian method with a Markov Chain Monte Carlo (MCMC) process to estimate posterior distributions of parameters. This allows us to use posterior densities of ω to further investigate whether dN/dS is greater at any particular codon in one dataset vs. the other without the need for nested models. This can only be done if datasets are the same length, encode for homologous genes, and have reliable alignments of codon positions. By sampling from the distribution of ω values we are able to determine the ratio of ω from *Physalis* (ω_1 relative to *Solanum* (ω_2. Rejection of the null hypothesis that sites have equivalent ω values is observed when the 95%

posterior density of ratios includes 1 (H_0: $\omega_1/\omega_2 = 1$). Posterior probabilities are then estimated using the new ω_1/ω_2 posterior density.

Rather than estimating ω for each dataset using a variable model along pre-defined blocks of codons (Wilson and McVean 2006; Vieira et al. 2007), we assumed an independent model for each site with an improper inverse distribution of rates. The MCMC chain was iterated over 500,000 generations sampling every 100th generation. We ran each dataset twice to check for convergence and removed a burn in of 50,000 generations using R (http://www.r-project.org/). The chain generates upper and lower posterior densities (highest posterior density HPD) to determine mean point estimates of ω at each codon position for each dataset. Because the independent model is computationally intensive, we ran the OmegaMap analyses using the Cornell BioHPC server (http://cbsuapps.tc.cornell.edu/omegamap.aspx). The upper and lower HPD of ω 's from each dataset were then combined and re-sampled after a burn in of 25,000 generations to get HPD's and the geometric mean for the ratio of ω's using R.

Fixed Effects Likelihood

We also used a fixed-effects likelihood (FEL) method to infer differential selection at individual sites among datasets (Kosakovsky Pond and Frost 2005). This method was first used to test whether HIV strains are experiencing differential selection in unique host populations (Kosakovsky Pond et al. 2006). FEL differs from the REL type models of PAML and the coalescent method of OmegaMap in that dN and dS are estimated at individual sites directly rather than using pre-defined distributions of rates (Kosakovsky Pond et al 2009). The entire alignment of each dataset is first used to estimate global parameters such as nucleotide frequencies, topology, and branch lengths. These parameters are then fixed throughout the selection estimate procedure. The null model H_0: $dS_1/dN_1 = dS_2/dN_2$ and alternative model H_A: where dS_1, dN_1, dS_2, dN_2 are free to vary are fitted to every codon and because they are nested, standard p-values based on LRT's can be used to determine significantly different sites. We estimated selection using the *CompareSelectivePressure* batch file in HYPHY v0.99. Actual dN/dS values for each dataset were then checked for any potential false positive estimates of selection. Here, it is possible for the model to reject the null hypothesis that dN/dS are equivalent across datasets but may not actually have $\omega > 1$.

We conducted simulations for *Physalis* and *Solanum* datasets independently to determine the power of the FEL test for given *p*-values. We simulated 100 replicates of each dataset and corresponding phylogeny using the site by site rate estimates by the FEL method with 25% of sites evolving neutrally. This produced 13100 sites with non-zero rates (131 codons x 100 replicates) to estimate false positive rates over bins of *p*-values of width 0.01.

Substitution rates

A phylogenetic approach as described in Lemey et al. (2006) was used to estimate average relative dN and dS along terminal branches of *P. longifolia* and *S. chilense* and compared whether these rates have varied through time across taxa. Samples of Bayesian relaxed-clock phylogenies were first generated using a coalescent model with constant population size in Beast v.1.4 (Drummond and

Rambaut 2003). Each dataset was rooted with one outgroup S-RNase from *Antirrhinum hispanicum* (Ahis5). After generating a sample of 1000 ultrametric trees, a sub-sample of 500 trees was used to infer branch lengths in expected numbers of dN and dS rates separately for each topology. Upon fixing branch lengths for each tree, we used a maximum likelihood procedure in HYPHY (a script for this type of analysis can be found at: https://perswww.kuleuven.be/~u0036765/analyses/AbsolutedNdSRates/) to infer the codon model substitution parameters (MG94xHKY85) and expected numbers of dS and dN substitutions along each branch. Because the likelihood method is computationally intensive, we limited our comparative datasets to these two species because they are the most extensively sampled for S-RNase diversity and possess similar numbers of alleles. Here we are primarily concerned with terminal branch estimates for comparison of average dS and dN across taxa because these branches represent the extant variation upon which we are estimating dN/dS ratios. These are reported in terms of numbers of dN or dS substitutions per branch or lineage. We are thus able to determine whether dN/dS ratios are affected by overall lower synonymous substitutions in *Physalis* relative to *Solanum.*

Results

Sequences and Phylogeny Construction

The Bayesian consensus tree topology with S-alleles from the combined datasets is shown in Figure 1. The three restricted *Physalis* lineages (branches A, B and C in Figure 1) are consistent with previously published topologies (Igic et al.

2004; Stone and Pierce 2005; Paape et al. 2008) that use S-alleles from more genera and illustrate re-diversification from within those respective lineages. No *Solanum* alleles are found within those lineages. Estimates of average pairwise nucleotide diversity (π) show synonymous divergence is greater for both species of *Solanum* while non-synonymous divergence is similar among taxa (Table 1). A greater accumulation of synonymous substitutions is expected for *Solanum* S-alleles if these lineages are in fact older than those of *Physalis* as suggested by previous studies (Richman et al. 1996; Richman 2000; Paape et al. 2008).

Distribution of dN and dS Rates

A REL approach was used to compare the distributions of dN and dS across both datasets and found that they differed significantly in three of four LRT's (Table 2).

The alternative hypothesis (H_A) where the rates of dN and dS were free to vary had the highest log-likelihood score (lnL = -16749.63). The null model (a) that constrains both datasets to have equivalent dN/dS ratios was strongly rejected ($p < 0.0001$; df =1) while the null model (b) constraining the proportions of selected sites across datasets was not rejected ($p < 0.165$; df =1). This test allows dN/dS to vary freely but enforces the proportions (p_1 and p_2) in the positive selection class to be equal and thus closely resembles H_A. Because this test does not indicate specific sites that differ for dN/dS ratios, subsequent analyses (results below) were conducted to determine locations along the S-RNase sequence selection differs among genera. The selective regime (c) test was also strongly rejected ($p < 0.0001$; 1df) but it is unlikely due to variation in proportions of selected sites based on the results of (b) and is largely the result of dN/dS variation across datasets. The shared distributions test (d) combines the joint distributions of dN and dS for both datasets and was also found to have a significantly lower likelihood ($p < 0.0001$; 10 df) than H_A which allows for variation in rates in both datasets.

Paml 3.15

We estimated positive selection at individual sites using the Nielsen and Yang (1998) method implemented in PAML v3.15. These results showed considerably more positively selected codons in *Physalis* than *Solanum* indicated by posterior probabilities > 0.99 (Figures 3 and 4). Because we cannot determine whether these sites differ significantly between datasets under the current framework of the maximum likelihood method in PAML, we employed the Bayesian method described by Wilson and McVean (2006) to generate HPD's for point estimates of ω. We first compared our results from OmegaMap with the Nielsen and Yang M3 model for both datasets. Posterior probability scores show remarkably consistent trends across methods for each dataset, though some sites have higher scores using M3 in *Solanum*. Most importantly these results demonstrate that both methods are able to consistently identify similar sites under positive selection upon which to estimate ω values. Wilson and McVean (2006) suggested that inconsistencies between their coalescent method results for estimating ω and those of *codeml* were likely the result of recombination. Although we did not detect the presence of recombination in either dataset (*not shown*) using the likelihood permutation test described by McVean et al.

(2002), historical recombination events among S-alleles have been suggested (Wang et al. 2001; Vieira et al. 2003) but are assumed to be rare. It is also unclear how extreme hyper-variability among regions of codons and the assumed prior rate distributions (general discrete: PAML vs. uniform: OmegaMap) at sites may confound analyses in each method.

OmegaMap ratio of ω*'s*

Estimations of the mean and upper and lower highest posterior densities (HPD's) for ω were used from each dataset to generate distributions of the ratio of ω's from the individual OmegaMap results (Figure 4). Confidence intervals (HPD's) that do not cross 1 (dotted line in Figure 4) indicate that the point estimates of ω values from each dataset are significantly different. The HPD's of ω_1/ω_2 are more heavily concentrated in the upper half of Figure 4 indicating that *Physalis* has higher dN/dS ratios at more sites. Not all sites with significantly different ω_1/ω_2 ratios had posterior probability scores for dN/dS when analyzed as separate datasets. We therefore culled sites that showed ≥ 0.95 posterior probabilities for both OmegaMap and PAML analyses for either *Physalis* or *Solanum* (or both) where all but 3 sites had posterior scores ≥ 0.99. Figure 5 shows dN/dS for 36 sites that had significantly higher ω_1/ω_2 and posterior probabilities ≥ 0.99 for *Physalis*. This test indicates that positive selection is significantly stronger within the 3 diversifying lineages of *Physalis* S-alleles relative to those of *Solanum*. Hypervariable (HVa and HVb) and conserved regions (C3 and C4) correspond to those identified in previous studies that estimated selection (Ioerger 1991; Takebayashi et al. 2003; Savage and Miller 2006;

Vieira et al. 2007; Igic et al. 2007). We have termed the region between C3 and C4, V1 for ‘variable region 1’ and V2 for the variable region at the 3’ end of C4.

Fixed Effects Likelihood

The FEL method also predicts that several codons in *Physalis* are under significantly greater positive selection than *Solanum.* This method is clearly more conservative than the Bayesian method as only 16 sites were predicted to be differentially selected (Figure 6) at the $p \leq 0.05$ level of significance and one site with $p = 0.08$. All but six sites were also identified by the coalescent method (Table 3). Because this method does not utilize rate distributions across sites, it is sensitive to the number of taxa present in each dataset (Kosakovsky Pond et al. 2009). We performed a power test to determine whether p-values ≤ 0.05 were sensitive to potential type II errors for the FEL analysis. Figure 7 shows that the power to detect positively selected sites for *Physalis* is 39.4% and 34% for *Solanum* at $p = 0.05$ implying that the power to detect selection by this method is quite low for these data sets. It is important to clarify that the false positive rate for sites predicted under this method is also low, 4.3% and 4.9% for *Physalis* and *Solanum* respectively. This means that when is a site is predicted to be under selection, the probability that it is truly under selection is ≥ 95%. So while the method is not extremely powerful for our datasets, it nevertheless appears to have considerable accuracy for the sites it does predict to be under selection. The accumulation of synonymous substitutions (Table 1) in *Solanum* S-RNases undoubtedly contributes to lower overall dN/dS, likely a result of the their greater age. Although higher dN/dS in *Physalis* can also be caused by lower dS (and younger age), the FEL test shows that for significantly differentially selected sites, dN/dS is indeed higher for this genus (Figure 6) for relevant codons.

Substitution rates

We estimated average dN and dS for *P. longifolia* and *S. chilense* at terminal branches by sampling from ultra-metric Bayesian coalescent trees to test whether ω values may be influenced by increases or decreases in either type of substitution (Table 5). We used only terminal branches because they reflect extant variation and the most recent patterns of diversification in dN and dS rates along a phylogeny. Average non-synonymous substitution rates per site per branch are nearly double for *P. longifolia* Threlative to *S. chilense*. As expected, dS is ten-fold higher for *S. chilense* indicating the younger age of *P. longifolia* S-alleles.

Discussion

S-allele variation provides one of the most extensive within-locus polymorphisms by which to estimate positive selection comparatively across taxa. The tests applied here all indicate that *Physalis* S-RNases exhibit statistically greater levels of positive selection than *Solanum*. Our first test was a REL method that uses a series of likelihood ratio tests to compare distributions of dN, dS and dN/dS among two datasets. This method indicated significantly greater substitution values for *Physalis*, though not a significant difference in the proportion of codons in each rate class. Therefore this test fails to show evidence for an increased proportion of selected sites in *Physalis* relative to *Solanum* but does provide evidence for stronger selection in *Physalis*. Because the particular sites under selection, in each data set and the sites that are under stronger selection in *Physalis* are not indicated by this analysis. This test is in essence a generalization used to compare rate distributions among homologous or non-homologous sequences, not for comparing which sites may under differential selection in either dataset. We conducted further tests to compare which sites may be under stronger selection in each dataset.

Posterior probability scores for the commonly used general discrete model (M3) of Nielsen and Yang (1998) and the Bayesian coalescent method of Wilson and McVean (2006) suggest that these methods perform similarly on these datasets. Under the current framework PAML does not permit comparisons of posterior distributions of ω (dN/dS) values as does the Bayesian coalescent method, allowing one to assess the level of support for differences in the strength of selection on the same codon position in different data sets. This analysis found considerably stonger selection at several codons in *Physalis*. The phylogenetic branch-site method of Zhang et al. (2005) that is included in PAML is not appropriate for S-RNase datasets or other highly polymorphic systems under strong diversifying selection. The limitation of this method is that it assumes a null model of uniform selection among 'background' branches that is not biologically realistic (Kosakovsky Pond et al. 2009) because positive selection is clearly evident (though differing in magnitude) for both *Physalis* and *Solanum* S-alleles. By sampling from the MCMC distribution of ω values and calculating their ratios, we were able to simply test whether these values differed significantly at each site for two datasets using a coalescent rather than phylogenetic approach.

The FEL method was used to confirm whether similar sites could be identified using a maximum likelihood approach that does not rely on rate distributions as in the coalescent method (Wilson and McVean 2006) or the Yang and Nielsen (1998) M3 model. This method clearly appears more conservative than M3 or the coalescent methods. This is not surprising given that this method relies upon actual variation at a given codon rather than a cumulative distribution among adjacent sites. Kosakovsky Pond and Frost (2005) demonstrated this using simulated datasets of increasing sample size where REL methods were considerably more powerful at detecting selection at low *p*-values for datasets of 32 and 64 sequences. For example, their simulation using 64 sequences showed that REL had power to predict positive selection ≥ 90% of the time. Our simulation results do show low false positive rates at these *p*-values suggesting a high probability of correctly identified sites for those indicated, but at low *p*-values (0.01- 0.1) the power to detect positively selected sites was below 50% (Figure 7).

The coalescent and FEL methods both predict the greatest differences among the genera compared in the magnitude of positive selection in the previously identified hyper-variable regions HVa and HVb (Ioerger 1991). The hyper-variable regions are thought to play a major role in determining specificity (Ioerger 1991, Ishimizu et al. 1998). Matton et al. (1997; 1999) demonstrated alteration of specificity using mutagenesis experiments involving these hypervariable regions. These studies showed that as few as 4 amino acid changes in corresponding positions of the S_{11} and S_{13} S-RNases of *S. chacoense* could alter specificity to that of the alternative allele. However, entire domain swapping in studies by Kao and McCubbin

(1996) and Zurek et al. (1997) using S-RNases of *Petunia inflata* and *Nicotiana alata* respectively, suggest that while HVa and HVb are important, other regions are also likely involved in recognition at least in some alleles or lineages. Consistent with this idea, both methods also show considerable differential selection in the V2 region near the 3' end of the S-RNases. Savage and Miller (2006) and Miller et al. (2008) also show many sites under selection in this region in *Lycium* S-RNases. It is clear from M3 and OmegaMap analyses on individual datasets as well as the ω_1/ω_2 test that *Physalis* has many more sites under selection in all variable regions confirming the results from the REL analysis.

The role of selection at the amino acid level is important when considering the co-evolution of the pollen component of the S-locus. It is expected that corresponding stigmatic and pollen S-allele cognates exhibit similar levels of divergence because a mutation in one gene but not the other would result in loss of SI. Several molecular studies implicate proteins encoding F-Box genes as the pollen S gene (Ikeda et al., 2004; Sijacic et al 2004; Qiao et al., 2004a,b; Ushijima et al., 2004) that are cognate to stigmatic S-RNases in the major plant families with this type of SI system. While only a few polymorphic pollen S alleles have been identified in these studies, they show greatly reduced divergence relative to putatively corresponding S-RNases, at least in some taxa (Sassa et al. 2007; Newbigin et al. 2008). Co-phylogenies show that this pattern also varies for each species studied thus far (Newbigin et al 2008). If this pattern is consistent for total pollen S polymorphism, it will result in greatly reduced levels of dN/dS relative to cognate S-RNases and confound models of selection on co-evolving sites. Once samples of pollen S sequences are near S-RNase numbers, the strength of selection between cognates can be analyzed by dN-dS rate distribution REL method used here.

Studies of statistically based positive selection on codons have been understandably met with some criticism (Hughes 2007; 2008). To assume positive selection is occurring at each site one has to ask, what is the selective agent? A mutation to a new functional S-allele has a selective advantage inverse to its frequency in the population. This has been termed 'rare allele advantage' because pollen bearing this mutation has access to any mate in the population. This principle is similar to frequency dependent pathogen escape from immune system recognition, though it differs from S-alleles with respect to fitness consequences. Diversification among *Physalis* S-alleles was also likely driven by population size expansion following some sort of severe restriction, as perhaps resulted from long distance dispersal. The numbers of extant S-alleles indicate similar current effective population sizes (N_e) in *Physalis* and *Solanum* (Wright 1939; Richman et al. 1996). The contrast that we see in signatures of selection in *Physalis* is likely a reflection of the increased strength of frequency dependent selection following a bottleneck and subsequent growth in population size. After the bottleneck, the population would have had fewer than the equilibrium number of S-alleles increasing the probability of fixation of mutations that alter specificity. In addition, the younger age of alleles has allowed less accumulation of synonymous substitutions.

Within population or species polymorphisms undergoing diversification highlight the need for statistical methods to detect whether selection is different at the codon level between datasets. Viral genes such as influenza, hepatitis-C and HIV strains are ideal candidates to detect differential selection pressures because they are well sampled and often adapted to local populations and directly reflect patterns of within host immune escape (Moore et al. 2002; Kosakovsky Pond et al. 2006; Poon et al. 2007). Comparisons of annual outbreaks of avian and human flu viruses (Campitelli et al 2006; Suzuki 2006) show distinct patterns of sequence evolution across taxa both geographically and annually, reflective of changing selective pressures (Furguson et al. 2003; Blackburne et al. 2006). To our knowledge, the only study that statistically compared selection pressures across populations of viral genes was that of Kosakovsky Pond et al. (2006) who developed the nested FEL method (used in our study) to identify selected sites among HIV genes from distinct African regions.

Studies of MHC variation across populations of closely and distantly related taxa of animals are too numerous to list (but see reviews by Sommer 2005 and Piertney and Oliver 2006), but represent cases where statistical comparisons of datasets would be useful. For example, Richman et al. (2003) found reciprocally monophyletic clades among MHC alleles in closely related *Peromyscus* species to be both the result of a bottleneck and intragenic recombination. Variation in these taxa among species appeared to be limited to < 20 amino acids and it would be of interest to test whether specific clades have selection toward different residues. In this case the coalescent method used in the present study would be useful as it accounts for recombination while individual species' sequence numbers are too low for the FEL test. In the only study outside of Kosakovsky Pond et al. (2006) to statistically compare codon selection pressures as described here, Blais et al. (2008) used the distribution of rates and FEL methods to detect adaptive divergence in MHC alleles in sympatric Lake Malawi haplochromine cichlids. Many other examples are known that include local pathogen mediated selection (Sommer 2005) and assortative mating cues in cichlids (Plenderleith et al. 2005) and sticklebacks (Milinski et al. 2005) where MHC diversity may be distinct across populations with selection towards specific residues, but have not been estimated comparatively.

The complementary sex determiner (*csd*) locus in several *Apis* species exhibits high polymorphism among alleles from three species but variation is unexpectedly restricted to monophyletic clades (Hasselman et al 2008). Hasselman et al. (2008) found unusually high turnover within this balanced polymorphism attributed to reduced population sizes throughout honey bee histories. While excesses of non-synonymous substitutions have been estimated (Hasselman and Beye 2004), no study has yet employed maximum likelihood selection analyses on codons. Sequences of *csd* genes have been well sampled and statistical comparisons between type I and type II selective pressures between species could prove interesting given the known genealogical characteristics.

While this list of examples is not exhaustive, it is apparent there are many cases that could benefit from statistical inferences of selection for different codons across populations or species. Existing methods that could be improved in this area include a branch-site method that does not constrain background branches to be evolving neutrally but allows them have sites that vary freely. Also, an extension of any maximum likelihood phylogenetic method that can statistically compare ω values between datasets that estimates posterior probabilities or LRT’s that a site is different under the REL method. This model would be similar to the FEL approach described here and in Kosakovsky et al. (2006) but would allow cumulative distributions across sites (as with M3 and M8 models in PAML) to increase power when datasets possess fewer sequences, which is becoming more computationally feasible.

Acknowledgements

Thanks to Danny Wilson for R code to estimate ratio of omegas as well as valuable suggestions and assistance with software. Thanks also to Sergei Kosakovsky Pond for consultation regarding various methods and for support with simulations and HYPHY assistance. Support for this work was provided by NSF DEB-0108173 and DEB-0639984 to J.R.K.,

Table 1. Average pairwise nucleotide divergence π among S-alleles for each species and genus estimated using DNASP 4.0 (Rozas et al 2003). Synonymous divergence is elevated in *Solanum* S-alleles, presumably because of greater age.

Taxa (*n* alleles)	**Synonymous (πs)**	**Non-Synonymous (πn)**	**All Sites**
Physalis (47)	0.33	0.34	0.33
P. cinerascens (12)	0.37	0.37	0.37
P. longifolia (37)	0.33	0.33	0.33
Solanum (49)	0.48	0.34	0.37
S. carolinense (17)	0.47	0.35	0.38
S. chilense (32)	0.5	0.34	0.37

Table 2. Rate distributions of non-synonymous and synonomous substitutions in each each dataset. Null models (a-d) were tested using likelihood ratio tests (LRTs) against the alternative model where dN and dS rates are free to vary in each dataset. Significance of $p \leq 0.05$ was determined using χ^2.

H_A: Rates free to vary

Log likelihood: -16749.63 Parameters: 229

Inferred rates for *Physalis:*				Inferred rates for *Solanum:*			
dN/dS	dS	dN	Prob	dN/dS	dS	dN	Prob
2.663	1.047	2.788	0.463	1.139	0.942	1.073	0.353
1.000	0.814	0.814	0.311	1.000	2.000	2.000	0.094
0.000	0.580	0.000	0.081	0.496	0.800	0.397	0.274
0.177	1.487	0.262	0.144	0.083	0.933	0.077	0.279

a) H_0: Same strength of selection

Log likelihood: -16765.49 Parameters: 228

Inferred rates for *Physalis*:				Inferred rates for *Solanum*:			
dN/dS	dS	dN	Prob	dN/dS	dS	dN	Prob
1.664	1.261	2.099	0.466	1.664	0.781	1.300	0.337
1.000	0.647	0.647	0.312	1.000	2.241	2.241	0.086
0.000	0.470	0.000	0.081	0.527	0.884	0.466	0.290
0.171	1.222	0.209	0.141	0.087	1.001	0.087	0.286

Are selection strengths (dN/dS) different?

LRT = 31.722 **$p < 0.0001$**; DF = 1

b) H_0: Same proportion of selected sites

Log likelihood: -16750.60 Parameters: 228

Inferred rates for *Physalis*:				Inferred rates for *Solanum*:			
dN/dS	dS	dN	Prob	dN/dS	dS	dN	Prob
2.737	1.042	2.851	0.397	1.143	0.949	1.085	0.397
1.000	0.900	0.900	0.339	1.000	2.065	2.065	0.081
0.000	0.573	0.000	0.084	0.491	0.804	0.395	0.258
0.216	1.297	0.280	0.180	0.082	0.939	0.077	0.264

Are the proportions of codons under selection different?

LRT =1.929 $p < 0.165$; DF = 1

c)H_0: Same selective regime (dN/dS and proportions)

Log likelihood: -16766.96 Parameters: 228

Inferred rates for *Physalis*:				Inferred rates for *Solanum*:			
dN/dS	dS	dN	Prob	dN/dS	dS	dN	Prob
1.636	1.318	2.157	0.397	1.636	0.805	1.318	0.397
1.000	0.703	0.703	0.348	1.000	2.341	2.341	0.074
0.000	0.472	0.000	0.087	0.517	0.894	0.463	0.265
0.193	1.136	0.219	0.169	0.086	1.022	0.088	0.264

Are selective regimes (dN/dS and proportions) different?

LRT = 34.647 **$p < 0.0001$**; DF = 2

d) H_0: Shared distributions of rates

Log likelihood: -16764.30 Parameters: 219

Inferred joint rates:			
dN/dS	dS	dN	Prob
2.507	1.034	2.593	0.189
1.000	1.139	1.139	0.338
0.543	0.797	0.433	0.251
0.086	0.988	0.085	0.222

Are the distributions different?

LRT = 29.350 **$p < 0.001$**; DF = 10

Table 3. Sites predicted to be under differential positive selection using the Bayesian ratio of omegas (ω_1/ω_2) test and the Fixed effects likelihood (FEL) test. For the ω_1/ω_2 test codons had a posterior probability score ≥ 0.99 unless otherwise indicated in parentheses. For the FEL test all sites listed had a *p*-value ≤ 0.05 except site 26. Hyper-variable regions HVa and HVb and conserved regions C3 and C4 correspond to those identified by Takebayashi et al. (2003) and Savage and Miller (2006). We have termed two additional variable regions V1 and V2. Bold indicates codons identified by both methods to be differentially selected.

S-RNase Region (codon positions)					
HVa (1-35)	HVb (44-62)	C3 (63-68)	V1 (69-84)	C4 (85-93)	V2 (94-131)
ω_1/ω_2 Test					
7,**8**, 10, 11, **13**,14, 15, 26, 33	44, **46**, 49 52, **53**, 56 (0.97), 59	-	69, 71, **84**	**87** (0.98)	**96,** 101, 104, 110, 111, 112, 119 (0.96), 120 (0.96), 123, 124 (0.97), **125** **127** (0.95), **129, 131**
FEL Test					
8, 9,**13,** 24, 26 (0.08), 31	**46, 53**	-	**84**	**87**	**96**, 99, 116, **125** **127, 129, 131**

Table 4. Estimates of average non-synonymous (dN) and synonymous (dS) substitutions at terminal branches (extant S-alleles) indicate that both rates are significantly different between species. Most importantly, dS is an order of magnitude greater for *S. chilense* indicating the younger age of *P. longifolia* S-alleles. These results suggest that while substitution/ mutation rates are similar for both taxa, the recent diversification of *P. longifolia* S-alleles (terminal branches) is indicated by lower dS.

Species (n alleles)	**dN/site/terminal branch**	**dS/site/terminal branch**
P. longifolia (37)	3.7 x 10^{-2} 1.48 x 10^{-3}	**5.7 x 10^{-3}** ± 2.06 x 10^{-4}
S. chilense (32)	2.0 x 10^{-2} ± 4.91 x 10^{-4}	**1.25 ± 10^{-2}** 3.01 x 10^{-3}

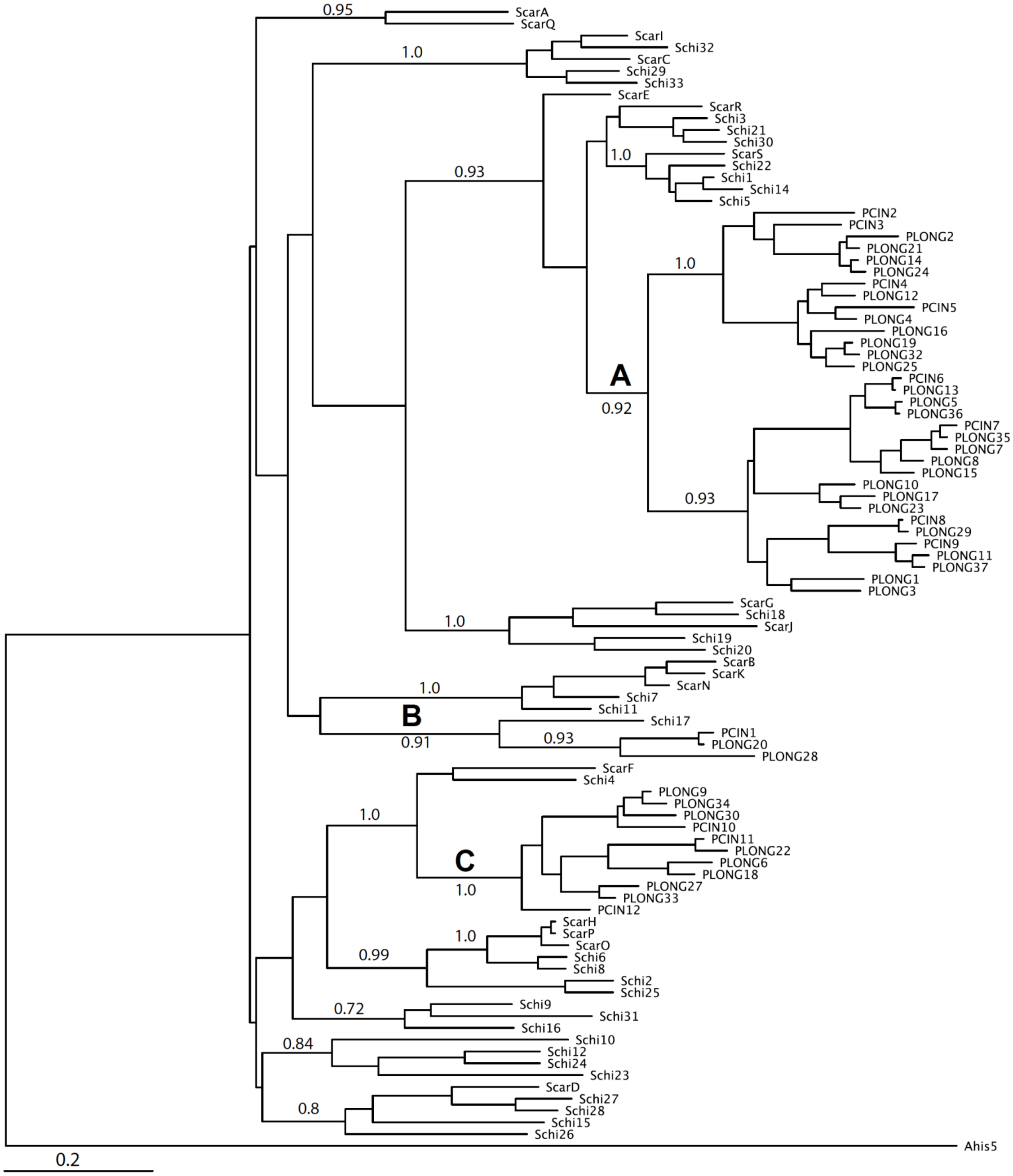

Figure 1. Phylogeny of *Physalis cinerascens* (Pcin), *P. longifolia* (Plong), *Solanum carolenense* (Scar), and S. *chilense* (Schi) S-RNases. Posterior probability scores show branch support for lineages of interest. The restricted (bottlenecked) lineages of *Physalis* are indicated at branches A, B and C. The phylogeny was created using Mr. Bayes v3.1 (Ronquist and Huelsenbeck 2003).

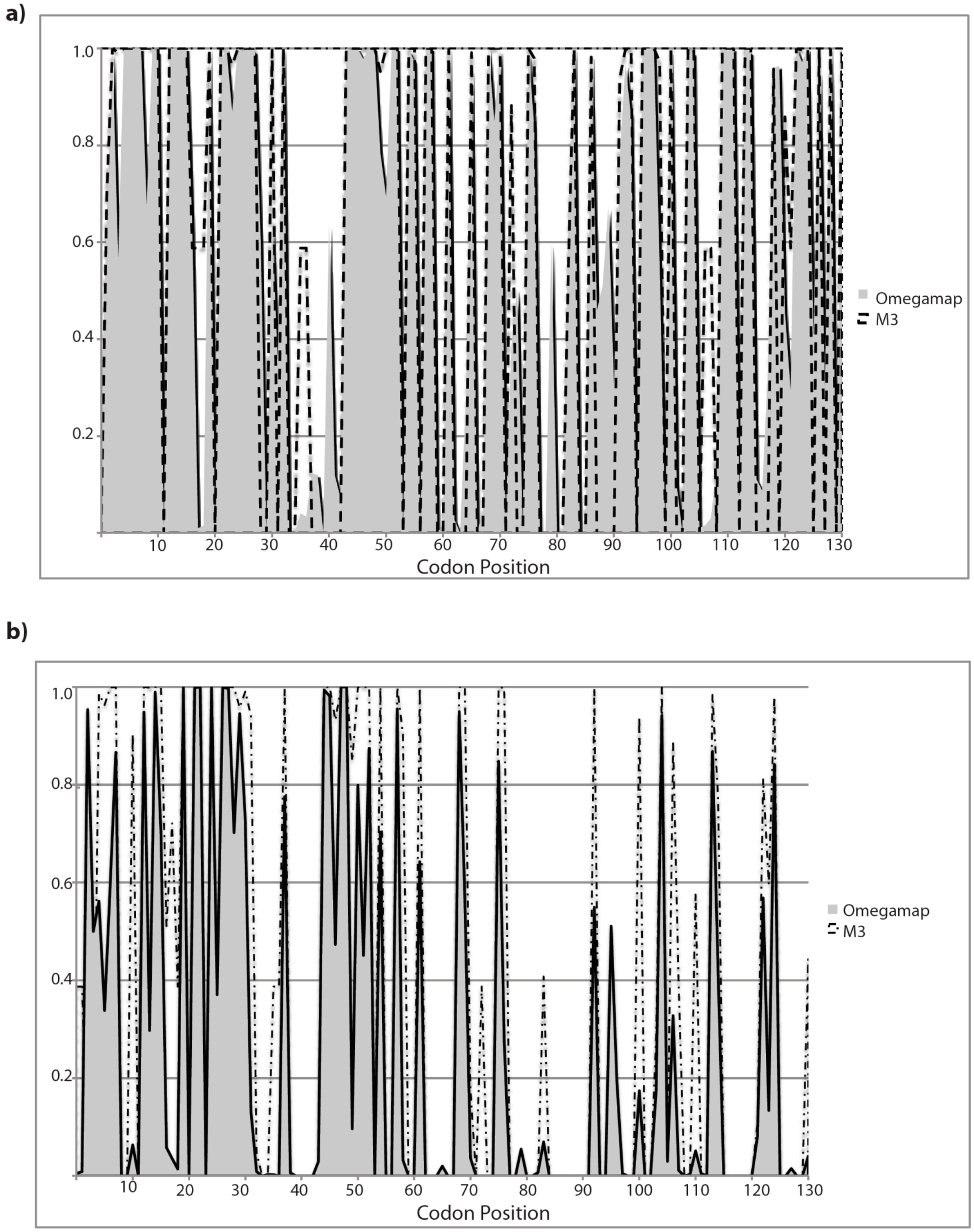

Figure 2. Posterior probability scores of sites predicted to be under positive selection in a) *Physalis* and b) *Solanum* using OmegaMap (Wilson and McVean 2006) (gray) and the general discrete model M3 (dased lines) of Nielsen and Yang (1998).

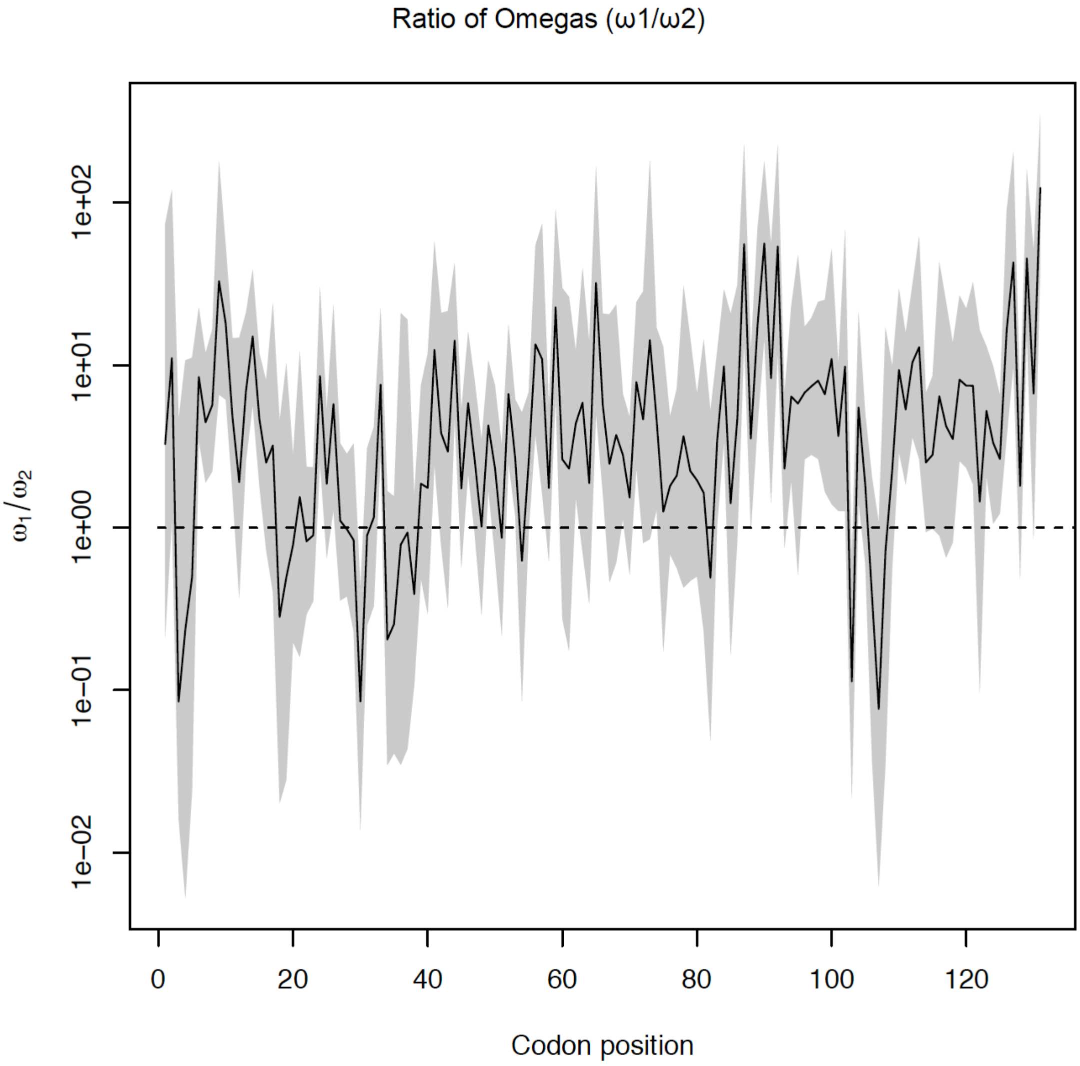

Figure 3. Bayesian estimate of the ratio of omega values (ω_1 = *Physalis* d_N/d_S; ω_2 =*Solanum* $d_N d_S$)foreach codon position. The gray region is the 95% highest posterior density (HPD) and the solid line is the mean of the ratios. If the HPD crosses the value 1 (dashed line) then the ratios are not significantly different. HPD's above the line indicate a higher ωfor *Physalis* than *Solanum* S-alleles.

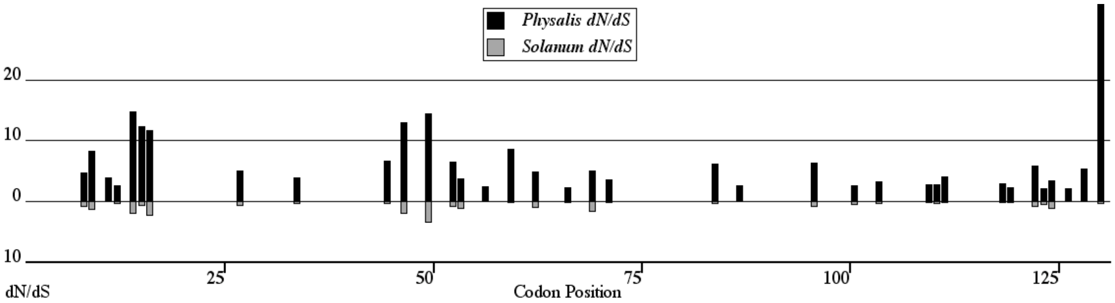

Figure 4. Contrast of point estimates of dN/dS for *Physalis* and *Solanum* for sites that were found to have omega ratios (ω_1/ω_2) significantly above 1 (from Figure 1). Sites indicated were first determined to be positively selected in at least one dataset based on posterior probability scores > 0.95 or 0.99 for both PAML and OmegaMap. For all sites, *Physalis* has higher dN/dS.

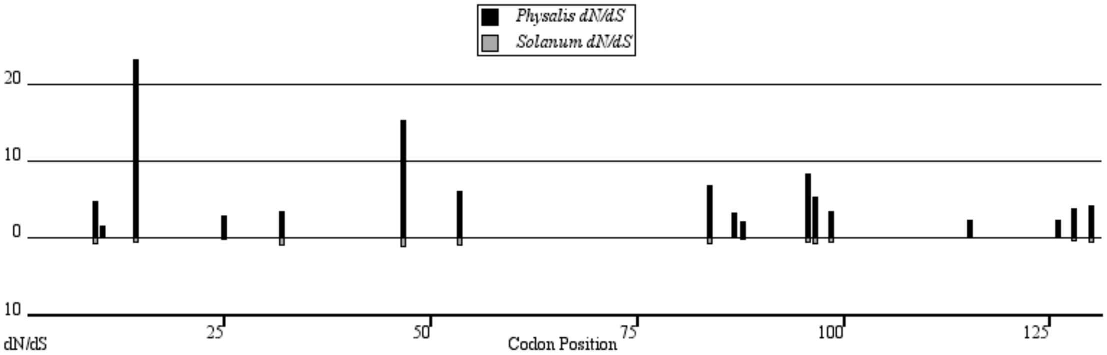

Figure 5. Fixed effects likelihood (FEL) comparisons of non-synonymous (dN) substitutions at sites predicted to be under significantly different selection pressure ($p \leq 0.05$). A total of 17 sites were determined to be greater for *Physalis* than *Solanum*.

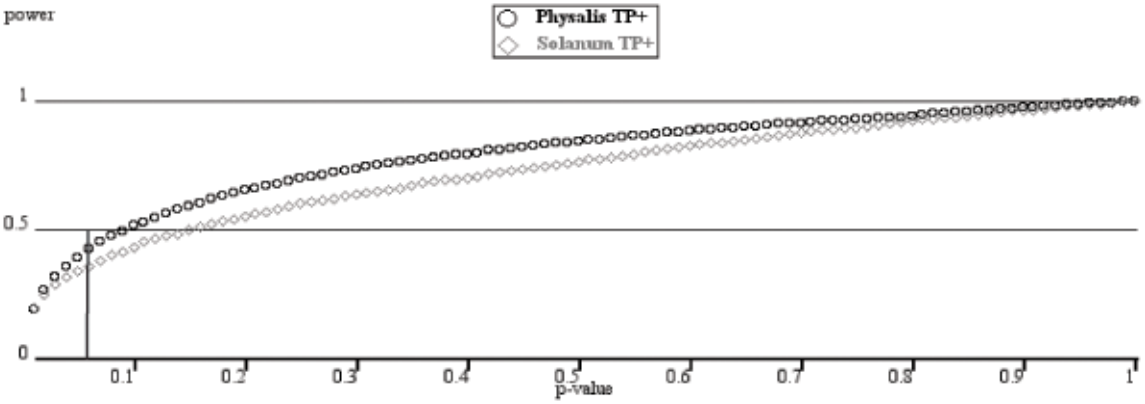

Figure 6. Power test for Fixed Effects Likelihood (FEL) test of positive selection on *Physalis* (n = 48) and *Solanum* (n = 49) datasets. For *p*-values of 0.05 (vertical line) positive selection is detected 39.3% for *Physalis* and 34% for *Solanum*. These simulations indicate that this method has low power to detect selection at p-values < 0.1. Data points represent true positive values (TP+) for 100 replicates**.**

References

Bechsgaard, J.S., V. Castric, D. Charlesworth,_ X.Vekemans, and M.H. Schierup. 2006. The transition to self-Compatibility in *Arabidopsis thaliana* and evolution within S-haplotypes over 10 Myr. Mol. Biol. Evol. 23:1741–1750.

Blackburne BP, Hay AJ, Goldstein RA. 2008. Changing Selective Pressure during Antigenic Changes in Human Influenza H3. PLoS Pathog 4(5): e1000058. doi:10.1371/journal.ppat.1000058

Blais J, Rico C, van Oosterhout C, Cable J, Turner GF, Bernatchez L. MHC adaptive divergence between closely related and sympatric African cichlids. PloS ONE. 2007

Campitelli L, Ciccozzi M, Salemi M, Taglia F, Boros S, Donatelli I, Rezza G. 2006. H5N1 influenza virus evolution: a comparison of different epidemics in birds and humans (1997-2004). J Gen Virol 87:955–60.

Clark AG. 1993. Evolutionary inferences from molecular characterization of self-incompatibility alleles. In: Takahata N, Clark AG (eds) Mechanisms of Molecular Evolution. Sinauer: Sunderland, MA, pp 79–108.

Drummond A, Rambaut A. 2003. BEAST version 1.3. Available: http://evolve.zoo.ox.ac.uk/beast.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:699-710.

Ferguson, NM, AP Galvani, RM. Bush. 2003. Ecological and immunological determinants of influenza evolution. Nature 422: 428-433.

Hasselman M, Beye M. 2004. Signatures of selection among sex-determining alleles of the honey bee. Proc. Nat. Acad. Sci 101: 4888-4893.

Hasselmann, M., X.Vekemans, J.Pflugfelder, N. Koeniger, G. Koeniger, S. Tingek, and Martin Beye. 2008. Evidence for Convergent Nucleotide Evolution and High Allelic Turnover Rates at the complementary sex determiner Gene of Western and Asian Honeybees. Mol. Biol. Evol. 25: 696–708.

Igic B, Bohs L, Kohn JR. 2004. Historical inferences from the self-incompatibility locus. New Phytologist. 161:97-105.

Igic B., W.A. Smith, K. Robertson, B.A. Schaal, and J.R. Kohn. 2007. The population genetics of the self-incompatibility polymorphism in wild tomatoes: I. S-RNase diversity in *Solanum chilense* (Dun.) Reiche (Solanaceae). Heredity 99:553-561

Ioerger TR, Gohlke JR, Xu B, Kao T-h. 1991. Primary structural features of the self-incompatibility protein in Solanaceae. Sex. Plant Reprod. 4:81–87

Ishimizu, T, T Endo, Y Yamaguchi-Kabata, KT Nakamura, F Sakiyama, S Norioka. 1998. Identification of regions in which positive selection may operate in S-RNase of Rosaceae: Implication for S-allele-specific recognition sites in S-RNase. FEBS Lett 440: 337-342

Kao, T.-h., and McCubbin, A. 1997. Molecular and biochemical bases of gametophytic self-incompatibility in Solanaceae. Plant Physiol. Biochem. 35:171-176

Kusaba, M., T. Nishio, Y. Satta, K. Hinata, and D. Ockendon. 1997. Striking sequence similarity in inter- and intra-specific comparisons of class I *SLG* alleles from *Brassica oleracea* and *Brassica campestris*: Implications for the evolution and recognition mechanism. PNAS. 94: 7673-7678

Kosakovsky Pond, S.L, S. D. W. Frost and S.V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208–1222.

Kosakovsky Pond SL, Frost SDW, Grossman Z, Gravenor MB, Richman DD, et al. 2006. Adaptation to different human populations by HIV-1 revealed by codon-based
analyses. PLoS Comput Biol 2(6): e62

Kosakovsky Pond, S.L Art F.Y. Poon, and Simon D.W. Frost. 2009. Estimating selection pressures on alignments of coding sequences: Analyses using *HyPhy*. The Phylogenetic Handbook. Cambridge University Press. in press

Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, et al. 2007. Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. PLoS Comput Biol 3(2): e29

Levin RA, Miller JS. 2005. Relationships within tribe Lycieae (Solanaceae): paraphyly of *Lycium* and multiple origins of gender dimorphism. Am. J. Bot. 92:2044–2053.

Levin RA, Shak JR, Miller JS, Bernardello G, Venter AM. 2007. Evolutionary relationships in tribe Lycieae (Solanaceae). Acta Hort. 745:225–239.

Lu, Y. 2001. Roles of lineage sorting and phylogenetic relationship in the genetic diversity at the self-incompatibility locus of Solanaceae. Heredity 86: 195–205.

Matton DP, Luu DT, Qin X, Laublin G, O'Brien M, Maes O, Morse D, Cappadocia M. 1999. Production of an S-RNase with dual specificity suggests a novel hypothesis for the generation of new S alleles. The Plant Cell 11: 2087–2097

Matton DP, Maes O, Laublin G, Qin X, Bertrand C, Morse D, Cappadocia M. 1997. Hypervariable domains of self-incompatibility RNases mediate allele-specific pollen recognition. The Plant Cell 9: 1757–1766.

McClure B.A., Haring V., Ebert P.R., Anderson M.A., Simpson R.J., Sakiyama F., and Clarke A.E. 1989. Style self-incompatibility gene products of *Nicotiana alata* are ribonucleases. Nature 342: 955-957.

McClure B.A., Gray J.E., Anderson M.A., and Clarke A.E. 1990. Self-incompatibility in Nicotiana alata involves degradation of pollen rRNA. Nature 347: 757-760.

McClure, B.A. 2004. *S-RNase* and *SLF* determine *S*-haplotype–specific pollen recognition and rejection. Plant Cell 16, 2840-2847.

McClure, B.A. 2006. New views of S-RNase-based self-incompatibility. Curr. Op. Plant Biol. 9, 639-646.

McVean, G. A. T., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241

Milinski M, Griffiths S, Wegner KM, Reusch TBH, Haas-Assenbaum A, et al. 2005. Mate choice decisions of stickleback females predictably modified by MHC peptide ligands. Proc Natl Acad Sci USA 102: 4414–4418.

Miller JS, Levin RA, Feliciano NM. 2008. A tale of two continents: Baker's rule and the maintenance of self-incompatibility in *Lycium* (Solanaceae). Evolution. 62-5: 1052–1065

Moore CB, John M, James IR, Christiansen FT, Witt CS, et al. 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science 296: 1439–1443

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–36.

Nunes, M.D.S., Santos, R.A.M., Ferreira, S.M., Vieira, J., and Vieira, C.P. 2006. Variability patterns and positively selected sites at the gametophytic self-incompatibility pollen SFB gene in a wild self-incompatible *Prunus spinosa* (Rosaceae) population. New Phytol. 172, 577-587.

Paape, T**.,** B. Igic, S. Smith, R. Olmstead, L. Bohs, J.R. Kohn. 2008. A 15-Million-Year-Old Genetic Bottleneck at the S-locus of the Solanaceae. Mol. Biol. Evol. 25: 655-663

Plenderleith M, van Oosterhout C, Robinson RL, Turner GF. 2005. Female preference for conspecific males based on olfactory cues in a Lake Malawi cichlid fish. Biol Lett 1: 411–414.

Poon AFY, Kosakovsky Pond SL, Bennett P, Richman DD, Leigh Brown AJ, et al. 2007. Adaptation to Human Populations Is Revealed by Within-Host Polymorphisms in HIV-1 and Hepatitis C Virus. PLoS Pathog 3(3): e45. doi:10.1371/journal.ppat.0030045

Rambaut A, Charleston M. Oxford University; Oxford: 2001. Tree Edit. Phylogenetic Tree Editor v1.0 alpha 8.

Rambaut A. 2002. Se-Al: Sequence Alignment Editor. Available at http://evolve.zoo.ox.ac.uk/.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3.1: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Richman, AD, MK Uyenoyama and JR. Kohn. 1996. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science. 273: 1212-1216

Richman AD and Kohn JR. 1999. Self-incompatibility alleles from *Physalis*: implications for historical inference from balanced genetic polymorphisms. PNAS 96: 168–172.

Richman, A. 2000. Evolution of balanced genetic polymorphism. Molecular Ecology. 9: 1953-1963

Richman, AD., Herrera, LG, D Nash. 2003. Evolution of MHC Class II E{beta} Diversity Within the Genus *Peromyscus*. Genetics 2003 164: 289-297

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Savage, A. E., and J. S. Miller. 2006. Gametophytic self-incompatibility in *Lycium parishii* (Solanaceae): allelic diversity, genealogical structure, and patterns of molecular
evolution at the S-RNase locus. Heredity 96:434–444.

Stone JL and Pierce SE. 2005. Rapid recent radiation of S-RNase lineages in *Witheringia solanacea* (Solanaceae). Heredity 94:547–555.

Sonneveld, T, Robbins, TP, Bošković, R, Tobutt, KR. 2001. Cloning of six cherry self-
incompatibility alleles and development of allele-specific PCR detection. Theoretical and Applied Genetics.102: 1046-1055

Suyama, M., D. Torrents, and P. Bork. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609-W612.

Suzuki, Y. 2006. Natural selection on the influenza virus genome. Mol. Biol. Evol. 23 :1902–1911.

Takebayashi, N., P. B. Brewer, E. Newbigin, and M. K. Uyenoyama. 2003. Patterns of variation within self-incompatibility loci. Mol. Biol. Evol. 20:1778-1794.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res., 25, 4876-4882.

Vieira CP, Charlesworth D and J Vieira. 2003. Evidence for rare recombination at the gametophytic self-incompatibility locus. Heredity 91:262-267.

Vieira, J., Morales-Hojas, R., Santos, R.A.M., and Vieira, C.P. 2007. Different positively selected sites at the gametophytic self-incompatibility pistil S-RNase gene in the Solanaceae and Rosaceae (*Prunus*, *Pyrus*, and *Malus*). J. Mol. Evol. 65, 175-185.

Wang, X., AL Hughes, T. Tsukamoto, T. Ando, and TH. Kao. 2001. Evidence That Intragenic Recombination Contributes to Allelic Diversity of the S-RNase Gene at the Self-Incompatibility (*S*) Locus in *Petunia inflata*. Plant Physiology. 125: 1012-1022

Wilson DJ and McVean G. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics.172:1411–1425

Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol Direct. 2006;1:34–53

Wright, S. 1939. The distribution of self-sterility alleles in populations. Genetics 24: 538–552.

Yang Z. 2000. Phylogenetic analysis by maximum likelihood (PAML). London: University College.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–18.

Zhang J, R. Nielsen, and Z. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472–2479

Zurek, D., Mou, B., Beecher, B., and McClure, B. 1997. Exchanging sequence domains between S RNases from Nicotiana alata disrupts pollen recognition. Plant J. 11:797-808

Chapter III

Evolutionary Genetics of Self-incompatibility in Papaveraceae

Abstract

The discovery of various forms of single-locus homomorphic self-incompatibility among several highly divergent flowering plant families represents a remarkable example of convergent evolution. Many Papaveraceae possess a gametophytic self-incompatibility (SI) system not homologous to any other SI mechanism characterized at the molecular level. Four previously identified full length S-alleles from the genus *Papaver* indicated remarkable divergence among haploptypes. Surprisingly these studies failed in attempts to sequence additional S-alleles despite earlier crossing data suggesting that more than 60 unique S-alleles occur in *Papaver rhoeas* alone. We identified 87_unique putative S-allele sequences from the taxa *Argemone munita*, *Papaver mcconnellii*, *P. nudicuale*, *Platystemon californicus* and *Romneya coulteri* using 5' and 3' RACE techniques. Hand pollinations among full-sib progeny of *A. munita* and *P. californicus* indicate a strong correlation between genotype and incompatibility phenotype. However, we also often find more than 2 homologous S-like sequences in individuals of *A. munita* and *P. californicus* with certain sequences co-segregating. Divergence estimates within and among taxa show *Papaver* alleles to be the most variable while divergence among sequences from other Papaveraceae is much lower. Genealogical analysis indicates very little shared ancestral polymorphism. Using coalescent methods, we found statistical evidence for different levels of recombination between *A. munita* and *P.californicus* which might help explain differences in levels of divergence and phylogenetic structure observed among samples of putative S-alleles from each genus. Estimates of positive selection at individual codons appears sensitive to the

level of variation within the taxa used. Most sites appear to be under selective constraint with few sites undergoing positive selection, suggesting that self-recognition may depend on amino acid substitutions at only a few sites.

Introduction

Angiosperms are known to possess various genetic mechanisms to reduce inbreeding by rejecting self-pollen. Single locus gametophytic self-incompatibility (GSI) occurs in several plant families (deNettancourt 1977) and is defined by rejection of haploid pollen in the stigma or style if the pollen carries a self-incompatibility allele (S-allele) that matches either allele in of the ovule parent. Crossing studies (Lawrence and O'Donnell 1981) of self-incompatible *Papaver rhoeas* indicated that this species maintains an estimated 66 S-alleles that segregate at a single heterozygous gametophytic *S*-locus (Lane and Lawrence 1993). Subsequent molecular studies led to the cloning of two polymorphic stigmatic S-alleles (Foote *et al*. 1994), but despite considerable efforts to clone additional haplotypes, Kurup *et al.* (1998) were only able to sequence one additional S-allele from *P. rhoeas* (S8) and one from the congener *P. nudicaule* (*Sn1*). It is thought that a very high level of divergence among *Papaver* S-proteins confounded the isolation of more S-alleles by standard PCR and hybridization techniques (Kurup *et al.* 1998). Sequence analysis of these alleles and studies characterizing the molecular mechanism of pollen tube inhibition revealed that this system bears no homology to the S-RNase system of Solanaceae, Plantaginaceae and Rosaceae, or the sporophytic SI system (SSI) found in Brassicaceae (reviewed in Wheeler and Franklin-Tong 2001; Franklin-Tong and

Franklin 2003), systems from which considerably more sequence data has been collected.

The S-allele products secreted in the pistil of *Papaver* are small proteins (~15 kDa) that bind to corresponding pollen S-receptors when haplotypes match, initiating a receptor mediated response characteristic of programmed cell death (Thomas *et al*. 2003; Thomas and Franklin-Tong 2004; Wheeler *et al*. 2009). The *Papaver* stigmatic S-proteins possess four conserved cysteine residues and the predicted secondary structure is comprised of ß-strand motifs linked by seven hydrophilic loops (Walker *et al*. 1996; Kakeda *et al*. 1998). Mutagenesis experiments (Kakeda *et al.* 1998) showed that residues in loop 6 were essential for the inhibition reaction of incompatible pollen. Other known types of molecular recognition proteins (eg. S-RNases, SSI, and MHC citations Please Fix) have distinct hypervariable regions interrupted by conserved regions contrasting with *Papaver* S-alleles where variability is at the 5' end (~30 amino acids) and at scattered residues throughout the sequence. An early scan of the *Arabidopsis* genome revealed as many as 100 'S-protein homologues' (SPHs) present (Ride *et al*. 1999) while another SPH was described in tomato (Testa et al. 2002). This indicates that the *Papaver* S-proteins may belong to a larger family of proteins but no function beyond SI has yet been described.

Many other Papaveraceae are known to be SI (Beatty 1936; Cook 1962; Hannan 1981), though S-locus polymorphism has not been investigated outside of *Papaver.* The S-RNase and SSI systems have both been shown to possess extensive shared ancestral polymorphism across genera (Ioerger *et al*. 1990; Richman *et al.* 1996; Kusaba *et al*. 1997). This reflects balancing selection preserving variation at the self-incompatibility S-locus for millions of years (Bechsgaard *et al*. 2006; Paape *et al*. 2008) and makes it possible to infer demographic events that occurred both before and after the origin of extant species (Richman *et al*. 1996; Miller *et al*. 2008; Guo *et al*. 2009; Foxe *et al*. 2009)

Here we attempt to isolate homologous S-alleles from SI species of Papaveraceae, each possessing unique growth habits and biogeography. We focus primarily on wild species that are close relatives to *Papaver* considering the previous difficulty isolating other S-alleles from within the genus. *Papaver mcconnellii* is a mostly tetraploid perennial that grows on rocky scree in Alaska and the Yukon Territory (Solstad 2008). Three species native to California are also investigated: *Argemone munita*, *Platystemon californicus*, and *Romneya couleri*. Both *A. munita* and *R. coulteri* are tetraploid and while the former typically grows as a desert annual, the latter is perennial and often grows clonally. *P. californicus* is a diploid summer annual and the only member of its genus.

Our goals in this study were to a) amplify putative homologous S-alleles from these species based on known *Papave*r S-allele sequences, b) determine whether putative S-haplotypes predict the SI phenotype in greenhouse crosses, c) assess genealogical characteristics and patterns of variation in putative S-alleles within and among species, d) because preliminary genotyping and sequencing of putative S-allele PCR products gives us reason to suspect gene duplication and potential recombination and gene conversion among paralogous copies, we test whether recombination is detectable among our samples of sequences and e) estimate sites that may be undergoing positive selection using both phylogenetic (Yang 2000) and coalescent methods (Wilson and McVean 2006). The limited number of *Papaver* S-alleles acquired by Foote *et al*. (1994), Walker *et al*. (1996) and Kurup *et al*. (1998) was previously insufficient to estimate positively selected amino acids.

Material and Methods

Plant Material

Stigmatic tissue for *Platystemon californicus* was collected from randomly sampled individuals from three previously documented populations (Hannan 1981, and personal communication): Lake Cuyamaca State Park east of San Diego (N 32°58' 59.1", W 116° 33' 45.4") Hastings Natural History Reservation (N 36° 23' 07.1", W 121° 33' 16.6"), and Carmel Valley Road near Carmel, CA (N 36° 26' 34.5", W 121° 38' 53.7"). Seeds of *P. californicus* were also collected from the Hastings Natural History Reservation and were germinated for greenhouse crossing experiments. *Argemone munita* stigmatic tissue and seeds were collected near Valentine Eastern Sierra Reserve, Mammoth Lakes, CA and Emerson Oaks UC Reserve near Temecula, CA. Bulk seed of *A. munita* collected from wild populations from San Diego County was used for greenhouse crosses. Full-sib families of both *A. munita* and *P. californicus* were produced from genotyped parental plants. Seeds were sown individually in 1" plug trays in moist SunGro Professional mix. Germination rates are very low for both species and typically require a minimum of 45-90 day cold treatment (4°C) before any seedlings emerge. *A. munita* was grown from seedlings in a glasshouse with natural light until flowering when hand pollinations were conducted. After germination, *P. californicus* was grown to flowering under fluorescent lights at 14/10 hr day/night light regimes. *Romneya coulteri* stigmas and leaves were collected from known populations from Del Mar, CA, Otay Mesa near San Diego, and UCSD. *Papaver nudicaule* (Iceland Poppy) was obtained from a commercial seed packet. *Papaver mcconnellii* was collected from a steep rocky scree outcrop near Highway Pass in Central Alaska (approx. N 63° 28', W 150° 10'). Seeds were cold stratified (4°C with 6hr light) for one month to induce germination, and plants were grown in 1:1:1 perlite:vermiculite:coconut coir in a glasshouse with natural light.

For *A. munita* and *P. californicus*, hand pollinations among sets of full-sib offspring were used to test the relationship between putative S-locus genotype and incompatibility phenotype. Flowers of both species were bagged prior to opening to prevent pollen contamination. Replicate pollinations of three or more crosses between each reciprocal pair were attempted, though not always feasible due to limited flowers. Pollinations were made by collecting anthers using a forceps and depositing pollen on the stigma. Crosses were made among individuals with matching sets of sequences ('common allele' treatment) and those that differed at a single putative allele ('semi' treatment). Self and non-matching genotype crosses were conducted in similar fashion.

The finding of more than 2 S-like sequences from single plants (see results) presented the possibility of a multi-locus system. To test for this we conducted twelve 5 x 5 reciprocal diallel crosses (4 in *A. munita* and 8 in *P. californicus*) among full-sib offspring. In a single-locus system, only four mating types are expected among full sibs so diallel crosses among five individuals should always uncover at least one incompatible pair. In multi-locus systems, the number of mating types among full sibs is expected to be greater, 16 if there are two heterozygous loci which both act to confer phenotype. These array crosses were conducted in the absence of any prior genotypic data among the individuals used.

Molecular Techniques

Amplification of putative S-alleles were performed using reverse-transcriptase (RT-) PCR from total RNA extracted from stylar/stigmatic tissue. RNA was extracted using Trizol (Invitrogen Corp.) or the PureScript kit (Gentra Systems Inc., Minneapolis, MN). For *Papaver* sp., some sequences were amplified from genomic DNA extracted from leaf tissues with the CTAB method (Doyle and Doyle 1987) or with PUREGENE DNA purification kit (Gentra Systems Inc., Minneapolis, MN). The forward primers Prho1F: 5'ATG/AAY/MRR/MGR/GGN/AAY/GG and Prho4F: GTG/CGH/ATA/ATG/AAY/ARR/AGA/GG were designed from a conserved amino acid region based on one published *P. nudicaule* and three *P. rhoeas* sequences and anneal to a region approximately 90 bp downstream from 5' end. These were used in conjunction with an oligo-dT reverse primer (AUAP, Invitrogen Corp.) for amplification of 3' ends of putative S-alleles from cDNA. Nested PCR was performed for several individuals of *P. californicus* due to non-specific amplification using the forward and reverse degenerate primers Prho1F and PopR3: GCC/ASR/TRH/ADV/ASC/WGC/C respectively. Amplified PCR products were then sub-cloned into competent *E. coli* cells using the TOPO 2.1 TA cloning kit

(Invitrogen Corp.). Restriction fragment analysis was used to identify unique products from clones.

Because there is an approximately 80-90 bp region upstream from the forward degenerate primers (based on the 4 known *Papaver* sequences), 5' rapid amplification of cDNA ends (5' RACE) was be performed using the Roche 2nd Generation 5' RACE kit (Roche Diagnostics). RNA was first transcribed into cDNA using allele specific reverse primers obtained from 3' end sequence data. This technique used the enzyme terminal transferase to ligate a poly-A region to the 5' end of cDNA and then one or two nested PCR reactions using additional allele specific reverse primers. To obtain the full-length sequences from genomic DNA, genome walking was conducted. After a fragment of a putative S-allele was sequenced, a specific primer for the allele was designed. To obtain full length *Papaver* sequences, genomic DNA was fragmented with restriction digest, and an adapter was attached to the ends. PCR was conducted with an allele specific primer and a generic primer on the adapter. Allele specific primers were also designed for genotyping parents and progeny using genomic DNA when possible (Supplementary material).

To assess tissue specificity of expression of putative S-alleles, RNA was isolated from stigmas, leaves, stems and buds of unopened flowers. Total RNA was treated with Ambion Turbo DNase (Ambion Inc.) according to the protocol to remove genomic DNA. Extractions were tested for DNA contamination by performing allele specific PCR prior to reverse-transcribing to cDNA. Because it is not possible to completely remove all genomic DNA, we used allele specific forward primers and a reverse abridged universal amplification primer that anneals downstream of the poly-

A tail on samples suspected to have remaining DNA. The quality of cDNA was then tested by PCR using *actin* primers.

Sequence Analysis and Phylogenetics

Sequences were aligned with BioEdit v7.05 (Ibis Therapeutics, Inc.) using the known *P. rhoeas* S-alleles as a guide. A total of 64 sequences were used to construct the phylogeny including sequences with 5' and 3' ends and those with only 3' ends. A maximum likelihood (ML) phylogenetic tree was constructed using PAUP (Swafford 2002) under the GTR+I+G substitution model. ModelTest 3.7 (Posada and Crandall 1998) was used to determine base frequencies (A = 0.31421 C = 0.15369 G = 0.24482 T = 0.28728) and the appropriate substitution model. Branch support for ML tree was obtained by bootstrap analysis of 1000 replicates using the same base frequencies as above. Mr. BAYES (Heulsenbeck and Ronquist 2001) was also used to confirm the maximum likelihood topology and branch support. The Bayesian analysis was run using a GTR+I+G substitution model across sites for 1,000,000 generations, sampling every 100th tree for a total of 10,000 trees. A burn-in of 2500 trees was discarded and posterior probabilities were calculated on the remaining 7500 trees. No root was specified for either ML or Bayesian analysis but the *P. rhoeas* S1 and S8 alleles are found to form distant sister group to all other alleles. Since the ML and Bayesian methods produced similar topologies, only the ML tree is reported.

Estimating recombination

Sequences from *P. californicus* and *A. munita* differed in average levels of divergence and in phylogenetic structure suggesting the possibility that recombination rates differed among these taxa (see Results). We utilized population genetics and coalescent approaches to detect recombination and estimate population recombination rates of 3’ datasets (bp 100-425) of *A. munita* (28) and *P. californicus* (24) sequences (Awadalla and Charlesworth 1999; McVean *et al*. 2002; Vieira *et al*. 2003). 3’ sequences were used for this analysis because they were the most numerous. We first tested for a significant association between linkage disequilibrium and physical distance between polymorphic sites using *permute* in OmegaMap v0.5 (Wilson and McVean 2006). This method is a non-parametric permutation test of whether the correlation between linkage disequilibrium (LD) and physical distance is stronger than expected under the null hypothesis of no recombination described in McVean *et al*. (2002). The permutation test uses three measures of LD as follows: $r^2 = \frac{\left(p_{AB}p_{ab} - p_{Ab}p_{aB}\right)^2}{p_A\left(1-p_A\right)p_B\left(1-p_B\right)}$, $D' = \frac{D}{D_{\min}}$ when ≥ 0 or when $D \leq 0$ $D' = \frac{D}{D_{\min}}$, and a modified r^2 that uses sites that may not have all four genotypes termed ‘G4’ described by Meunier and Eyre-Walker (2001).

We then estimated the population recombination rate, $\rho = 4N_e r$ for *A. munita* and *P. californicus* 3’ datasets using a composite likelihood estimator in OmegaMap also described in McVean *et al*. (2002). Because we are primarily concerned with average relative estimates of ρ between *A. munita* and *P. californicus*, starting prior values were set to a constant rather than variable model along the sequences using an improper inverse distribution of ρ over a range of $4N_e r$ between 10^{-7} - 10^{3}. The MCMC chain was run for 500,000 generations sampling every 500th generation.

Estimates of sequence diversity and positive selection

Average pairwise nucleotide divergence was estimated for synonymous and non-synonymous sites individually and on all sites using DnaSP (Rozas *et al*. 2003). Because we were able to amplify 3’ ends more readily, sliding window estimates of interspecific divergence was estimated for 5’ and 3’ alignments independently. The 5’ dataset included 10 full-length *Papaver* S-alleles, 11 *P. californicus* and 11 *A. munita* sequences. The 3’ alignment possessed 24 *P. californicus*, 28 *A. munita* sequences and 10 *Papaver* sequences.

To estimate the ratio (ω) of non-synonymous (d_N) to synonymous (d_S) substitution rates at each amino acid position we used the program *codeml* in PAML 3.15 (Yang 2000). Values of $\omega < 1$ for individual codons indicates purifying selection while sites with $\omega = 1$ are considered neutral. Positive selection at the amino acid level is inferred when $\omega > 1$. A series of nested neutral and selection models first developed by Nielsen and Yang (1998) and Yang *et al*. (2000) use likelihood ratio tests (LRT) to determine the model that best fits the data. The null model M1 (neutral) constrains all sites to be either of class $\omega = 0$ or $\omega = 1$ while the alternative model M2 (selection) adds a third class in which ω is free to vary at individual sites. Model M3 (selection) assumes three discrete site classes ($\omega_0 < 1$, $\omega_1 = 1$ and $\omega_2 > 1$) with three corresponding proportions (p_0, p_1, p_2) estimated from the data. Two similar models, M7 (neutral) and the alternative M8 (selection) assume a beta distribution of rates among 11 site classes. Models are rejected by $-2\Delta \ln L$ ($\Delta \ln L$ = the difference in log likelihoods of the two models) where significance is determined by χ^2 distribution with the degrees of freedom (df) equal to the difference in the number of parameters between models. Sites estimated to be under positive selection are determined by an empirical Bayesian approach (Yang *et al*. 2005) where posterior probabilities are estimated from rates within each site class. Because individual species possess several polymorphic alleles, comparisons of d_N/ d_S ratios can be made within and between species and can also be estimated cumulatively (Nielsen and Yang 1998; Ikeda *et al*. 2004, Turner and Hoekstra 2006).

We also used OmegaMap to determine whether different estimates of selection on individual codons are achieved using a coalescent method with recombination. An independent model for ω at each site was used for 5' datasets as these were sufficiently small (22-32 sequences, without and with *Papaver* spp., respectively) to allow for this computationally intensive MCMC chain of 500,000 generations. An improper inverse prior distribution of ω was set for the independent model.

Results

Sequences amplified were highly polymorphic as expected for putative S-alleles showing the nearest match to the known S-alleles of *P. rhoeas* in NCBI nucleotide Blast queries. RT-PCR of stigmatic tissue using degenerate primers resulted in the amplification of six new full-length putative S-alleles from *P. mcconnellii* and one from *P. nudicuale.* For *Papaver* spp. only one allele was detected for each of the seven individuals tested. Twenty-eight complete sequences as well as 12 partial sequences lacking 5'-ends were amplified from *A. munita* (Figure 1). We amplified one unique sequence from 8 *A. munita* individuals and two sequences from15, individuals. Three unique sequences were amplified from each of two individuals. A total of 31 *P. californicus* sequences were obtained using 3' RACE and nested PCR (Supplementary Figure X) and 10 sequences were found using 5' RACE. For *P. californicus*, only one product was found for 11 individuals, two products were found for 10 individuals while three or four products were found for 5 individuals. Over the region sequenced, identical sequences were found in 12 pairs of individuals. Two alleles were found for all 12 *Romneya coulteri* sampled for a total of six unique sequences.

Hand pollinations —*Argemone munita*

We genotyped two *A. munita* full-sib families ('8-1' and '25-4') using allele specific primers for the 5' ends of putative S-allele sequences. Family 8-1, generated by crossing plant 8 (with alleles *S8a S8b*) with plant 1 (with alleles *S1a S1d*), while family 25-4 was generated by crossing plants 25 (*S25a S25b*) with 4 (*S4a S4d*). Genotyped sequences segregated in an allelic fashion in both F1 families. When crossing full-sibs, we expect that crosses among plants with only one shared allele will be partially compatible and will set fruit, whereas crosses between matching genotypes (both alleles match) will be fully incompatible and will not lead to fruit set. As expected, 100% of the 64 crosses among individuals possessing only one shared allele produced fruit, however, 22% of crosses among individuals with matching genotypes also set fruit (Table 1). Fruit set following crosses among individuals with matching genotypes appeared to be genotype-specific. Matching-genotype crosses from family 8-1 never set fruit while those from family 25-4 produced no fruit unless the plants possessed allele '*S25b*'. If *S25b* was present, 59% of matching-genotype crosses set fruit. 100% fruit set was also observed in nine crosses among individuals where one parent came from family 8-1 and the other from 25-4 and therefore the parents shared no putative S-alleles in common.

It should be noted that autogamous fruit production in plants not carrying the *S25b* allele was rarely observed among *A. munita* despite large deposits of self-pollen visible on stigmas. Some autogamous seed set occurred plant 25, the parent that carried the *S25b* allele as well as in 3 of 21 of its progeny. Occasional fruit set was also observed among un-genotyped bagged and hand self-pollinated individuals of *A. munita* (7 of 61) indicating either additional variation in the strength of SI or that allele *S25b* was present among these plants. Outcross hand pollination from non-sib donors nearly always set fruit (136 of 141 pollinations). None of the F2 offspring from crosses among siblings germinated, perhaps due to inbreeding depression in addition to low rates of germination observed in the lab. It was therefore impossible to test for expected segregation of parental sequences. Though the sequences found in the parents and offspring of these 2 families appear to be allelic, more than 2 homologous products were found in the male parent of family 8-1, as well as one other plant assayed for putative S-allele sequences. This product was designated *S8e*, but only alleles *S8a, S8b, S1a* and *S1d* were genotyped in progeny.

Hand pollinations – *Platystemon californicus*

Prior to genotyping the progeny from two full sib families (34-26 and 34-21), we identified two stigma-S sequences from each parent: *S1a* and *S1b* from parent 34, *S3a* and *S3b* from parent 26 and *S2d* and *S2e* from parent 21. However, in both crosses, we were surprised to find that some of the stigma S sequences failed to segregate like alleles in the F1s, but rather co-segregated as if they are linked. In family 34-26, sequences *S3a* and *S3b* (hereafter *S3a,b* where *a* and *b* are unique PCR products) were both found in 12 of 22 individuals and never found separately, ruling out that they are in fact separate alleles of a heterozygous S-locus, as had been assumed when the cross was performed. A third product *S3a* and was only observed in the absence of *S3a* and *S3b*. No homologous product linked to *S3c* was detected. Sequences *S1a* and *S1b* from the other parent in this cross were never found together in F1 progeny, characteristic of alleles segregating at a heterozygous S-locus.

Family 34-21 also possessed a similar co-segregating polymorphism among F1 progeny. The PCR products *S2d and S2e* (hereafter *S2d,e*) from the male parent of this family were found together in 13 of 22 individuals (Table 2). Offspring of this cross either possessed both *S2d* and *S2e* sequences or else no sequence that could be attributed to the female parent. We apparently failed to amplify the alternative allele in this case.

The combined data of hand pollinations within the two families showed 93% fruit set when parents had one matching allele, while 12% of crosses among plants with matching putative S-locus genotypes set fruit (Table 2). The majority of unexpected fruit set (10 of 23 flowers) was found in crosses within Family 34-26 where only one allele was found (*S1b*) from one parent prior to conducting the cross. Allele specific PCR of the *S3c* product was ambiguous due to its short length and low divergence from *S3a*, so it is unclear whether fruit set was due to failure to correctly genotype some individuals or actual leakiness in the presence of allele *S3c*.

Hannan (1981) found that individuals from Hastings were highly SI but did report that 1% of hand selfed ovules developed into seed. Curiously, fruit set following 22 hand self-pollinations was never observed (data not shown) nor was there any autogamous fruit set of bagged flowers despite copious autogamous self-pollen on stigmas. Germination of seeds from one cross among F1's with non-matching genotypes was sufficient to perform crosses among F2's. In Family 34-21 F2, crosses among offspring with the *S1a S1b* heterozygous genotype produced no fruit in (N=17, Table 2). In both *A. munita* and *P. californicus,* fruit set was significantly higher when non-matching rather than matching genotypes were crossed (Table 3).

The finding of 3 S-like sequences from single individuals presents the possibility of a multi-locus system in which more than four incompatible phenotypes might result from a cross. We always found incompatible crosses in diallel crosses among 5 ungenotyped full sibs. Three full-sib 5 x 5 crossing arrays of *A. munita* had at least one pair of individuals that were reciprocally incompatible (Supplementary tables). An additional array possessed one pair of plants where three pollinations in one direction failed to produce fruit but the pair was not crossed reciprocally. In *P. californicus*, Four arrays among full-sib family 34-26 and four among family 34-21 each possessed one pair of reciprocally incompatible plants. The finding of incompatible pairs in these small diallels of full sibs supports the hypothesis of single-locus GSI in these taxa, as is found in *Papaver*.

Tissue specific RT-PCR

We investigated the expression of putative *S*-alleles in various tissues by allele specific amplification of cDNA (Figure 5, 6). In *P. californicus*, we used an individual from the crossing family 34-21 (Table 2). Allele *S1b* showed high expression in the stigma as expected for a *S*-allele, with a faint band appearing in the leaf tissue (Figure 5). While allele *S2d* was only expressed in the stigma, allele *S*2*e*, a potential paralogue linked to *S2d*, appeared to show expression in unopened buds, stems and leaves as well as in the stigma. However, sequencing of this amplicon from non-stigmatic tissue types revealed that it was not *S2e* but a paralog not previously identified from the parental stigmas. The *A. munita,* sequences *S*4*a*, *S*4*d*, *S*8*a*, *S*8*b*, *S*25*a* and *S25d* showed no expression in leaves but appear to be highly expressed in stigmatic tissue (Figure 6). Other tissue types were not tested for this species.

Phylogeny

The phylogenetic reconstruction of sequences from *Papaver, Argemone, Platystemon* and *Romneya* shows interesting and rather unexpected features (Figure 2). These previously isolated S-alleles show no shared ancestral polymorphism with the non-*Papaver* sequences (Figure 2). While the newly identified *P. mcconnellii* and *P. nudicaule* alleles show some trans-specific polymorphism with both *P. rhoeas* and *P. nudicuale,* they nevertheless cluster together with congeneric sequences rather than with sequences derived from other genera.

Sequences from *P. californicus* form a distinct, well-supported clade and show no evidence of shared polymorphism with other taxa (Figure 2). Paralogous sequences *S2d* and *S2e* from the two crossing families are found clustered in separate sub-clades within the *P. californicus* clade as do the linked homologs *S3a* and S3b. *A. munita, R. coulteri* and *R. coulteri* show some possible shared polymorphism but these lineages are not well resolved. Overall, there is little phylogenetic structure among sequences from *A. munita* compared to those from *P. californicus*(Figure 2) .

Sequence variablity

Average pairwise (π) estimates of synonymous, non-synonymous and total nucleotide divergence show that the previously sequenced S-alleles from *Papaver* are the most variable among the taxa studied (Table 5). Average divergence from *Papaver* is reduced slightly when sequences from *P. mcconnellii* are included due to the clustering of several alleles from this species with the previously published allele from *Papaver nudicaule* (Fig. 3). Pairwise divergence among *P. californicus* sequences appears higher than among *Argemone* sequences (Table 4). Sliding window analysis of 5' alignments show substantial variation in divergence among alleles from different taxa with *A. munita* sequences being the least variable as expected from π values (Table 5). Analysis of 3' datasets shows consistent variation at similar sites among taxa but is most pronounced among *Papaver* and *P. californicus* sequences throughout. These analyses confirm that variation is not found in distinct hypervariable blocks but dispersed throughout the protein at individual codons. Divergence is greatest among the known *Papaver* S-alleles from Kakeda *et al*. (1998; Table 4).

Recombination estimation

The finding of multiple homologous copies of S-like genes suggests the potential for recombination and gene conversion. We found significant evidence of recombination in *P. californicus*, but not *A. munita* (Table 5). However, the different outcomes of the analysis may be due to higher sequence diversity, and therefore higher power to detect recombination in *P. californicus*. Because the power to reject the null hypothesis of no recombination may be affected by θ, we conducted coalescent simulations with Hudson's (2002) ms and SeqGen (Rambaut and Grassly 1996) that held the ratio of ρ/θ constant but varied over ranges of θ (0.5-3.0). We set ρ/θ to the value estimated from *A. munita* (3.0; Table 6). Sequence length (327bp) and sample sizes (24) were set to match *A. munita*. Then the permutation test with r^2 was applied to simulated data and we calculated the frequency of null hypothesis rejection (Table 7). We found that for lower values of θ the power of the r^2 test for LD, the test that appears to be the most powerful (Table 7), decreases substantially (Table 8). Thus the value of θ (0.94; Table 6) for *A. munita* may be too low to supply adequate power to detect recombination by this means while the value of θ for *P. californicus* is adequate (Tables 6 and 7).

Despite the lack of statistically significant evidence of recombination among A. munita sequences, the estimated ratio (r/μ) of recombination rate (r) to synonymous mutation rate (μ) was higher for *A. munita* than for *P. californicus*

(Table 6). Comparisons of population recombination rate estimates (ρ) show that recombination in *A. munita* is significantly greater than for *P. californicus* (Table 6). This analysis also confirms that synonymous genetic diversity (**θ**) is significantly lower for *A. munita* than for *P. californicus*.

Estimates of positive selection

The putative S-allele sequences appear to be under the positive selection. Likelihood ratio tests show that models with positive selection fit the observed data better than neutral models (Table 5). The statistical significance holds for the genera with lower sequence diversity even after the Papaver spp. sequences are removed (Table 5b). The empirical Bayesian method (Nielsen and Yang 1998) identified only a few sites with statistically significant evidence for positive selection, while the majority of sites are found to be under purifying selection (Table 5). Separate 5' datasets including and excluding *Papaver* sequences were analyzed to determine their effects on estimates of selection in *A. munita* and *P. californicus*. Posterior probability scores indicate sites 53, 107 and 119 are under selection under models M2 and M8. The discrete M3 model shows no sites under positive selection when *Papaver* is present but that 6 sites (sites 53, 77, 83, 103, 107, and 119) have high posterior scores when those sequences are excluded. Model M8 shows higher mean dN/dS when *Papaver* is excluded suggesting those highly divergent sequences may have a dampening effect on estimating sites under selection.

We also estimated selection using OmegaMap because the coalescent methods indicate that recombination may be a factor. This method finds more sites under selection overall (Figure 7). Codons 13, 33 and 68 were predicted to be under positive selection with OmegaMap but not PAML. Codons 103, 107 and 119 were consistently predicted using both methods but site 53, which was identified by PAML, was not predicted using OmegaMap unless *Papaver* spp.were removed (Site 53: Posterior probability = 0.92). It appears that the independent model of rate variation may suffer from lack of power when only *A. munita* and *P. californicus* are included (22 sequences) as fewer sites are identified than when *Papaver* is included.

Discussion

In all Papaveraceae species studied here, we amplified polymorphic sequences that resemble the four previously described S-allele sequences from *Papaver rhoeas* and *P. nudicaule* (Foote *et al.* 1994; Kurup *et al.* 1998). Crossing experiments show that genotype significantly predicts fruit set in full-sib families of both *Argemone munita* and *Platystemon californicus*. In both species, crosses between parents with at least one non-matching allele nearly always set fruit, while crosses between individuals with matching putative S-locus genotypes displayed much lower rates of successful fruit set. In *A. munita*, fruit set following crosses among individuals with matching genotypes only occurred when allele *S25b* was present, suggesting that populations of this species may carry a mixture of functional and non-functional S-alleles. In *P. californicus*, fruit set following crosses among individuals with matching genotypes was not associated with any particular allele, but occurred with far lower frequency than following crosses among parents where at least one non-matching allele. In sum, putative alleles identified here appear to either be functional as part of the self-incompatibility system, or linked to the genes that are.

While two homologous products were found for several individuals of *A. munita* and *P. californicus* as expected of an obligately heterozygous GSI locus, only one product was found for several individuals of these species and for all individuals of *P. mcconnellii.* In addition, more than two sequences were found in some individuals of both *A. munita* and *P. californicus.* The finding of only one S-like sequence in some plants may imply that amplification is inconsistent among genotypes, perhaps due to PCR competition among alleles or that other lineages of S-alleles exist that are too divergent to be detected using our degenerate forward or reverse primers. Despite crossing data to show that *P. rhoeas* possesses more than 60 S-alleles (Lane and Lawrence 1993), Kurup *et al*. (1998) were only able to sequence two alleles from *Papaver* using various hybridization and PCR techniques in addition to the two previously described by Foote *et al*. (1994). Amplification of less than two alleles from individuals is common in natural population studies of other SI systems (Vieira and Charlesworth 2002; Mable *et al*. 2003) suggesting that amplification may fail, even when there is much prior knowledge of sequence variation.

More than two putative S-sequences in *A. munita* were found in 2 of 26 individuals genotyped. Finding more than two S-like sequences in at least some individuals was not surprising considering that this species is tetraploid. While polyploidy, or even duplication of just the pollen-determining part of the S-locus causes loss of self-incompatibility in RNase-based SI of Solanaceae and Plantaginaceae, (Golz *et al*. 1999, Sijacic *et al*. 2004; Qaio *et al*. 2004), this finding cannot be generalized to systems of incompatibility using other molecular mechanisms. In a broad survey, Mable (2004) found no significant association between ploidy level and incompatibility status and SI occurred in many polyploid species. Relevant to the situation described for *A. munita* is the case of the S-RNase system of the tetraploid *Prunus cerasus* (Rosaceae) where the species possesses both SI and SC individuals. Individuals that are SC possess haplotypes with either a loss of function mutation in the stigmatic S-RNase (Yamane *et al*. 2003) or the pollen-S F-Box gene (Hauck *et al*. 2006). Because it is not known whether the putative S-alleles from *A. munita* are in fact the actual genes responsible for stigmatic recognition or simply linked to the locus, either scenario for haplotypes possessing *S25b* are possible.

The full-sib progeny of both *P. californicus* families studied here showed two products from single parents (*S2d,e* and *S3a,b*) where individual offspring received either both sequences, or neither sequence, as if the two sequences are linked, rather than allelic. A third sequence (*S3c*) that appeared to be allelic to the co-transmitted sequences (*S3a,b*) was found in one family (34-26) but no sequence that appeared allelic was detected in the other family (34-21). The finding of multiple homologous copies of S-like genes was less expected for *P. californicus* as this species is diploid. In contrast to *A. munita,* fruit set in *P. californicus* was never observed following hand selfing or autogamous self-pollination. The majority of leakiness among individuals with matching genotypes in this species is found in crosses within family 34-26 where only one product (*S1b*) was detected. Because no second allele was detected in these plants, we cannot know whether fruit set resulted from leaky incompatibility or from failure to properly genotype some individuals.

The S-locus is subject to strong balancing selection because rare alleles are rejected by fewer potential mates than are common ones. The expectation for loci under balancing selection is that polymorphism will be preserved over very long time periods. However, our phylogenetic analysis indicates a general clustering of putative S-alleles according to genera, and the lack of trans-generic lineages with the only minor exceptions being among poorly supported nodes subtending alleles from *A. munita* and *R. coulteri*. While shared ancestral polymorphism often is observed in SI systems, examples exist where it is greatly reduced. Castric and Vekemans (2004) highlight a nearly monophyletic clade of *Brassica olearcea* S-receptor kinase (*SRK*; the stylar S-product) alleles with only two trans-specific lineages shared with *Arabidopsis lyrata* and none with *Arabdopsis halleri*. This was explained by the 15-20 million year divergence of *B. oleracea* from *Arabidopsis* (Kusaba *et al*. 2001) resulting in lineage sorting through time. However, lineage sorting should not result in the marked differences in average divergences among sequence seen when comparing *Papaver* sequences to all others.

Gene duplication and interlocus recombination could explain the phylogenetic patterns observed in this study. In *A. munita*, duplication may be a consequence of polyploidy. In *Platystemon* paralagous gene duplication may have occurred after divergence from *Papaver*. Gene duplication in *Platystemon* appears independent of *Papaver* as no linked or homologous S-like product has been detected in the genus (V.E. Franklin-Tong *personal communication*), despite considerable efforts to characterize the S-locus in *P. rhoeas* (Bosch and Franklin-Tong 2008). RT-PCR of *P. nudicaule* stigmas using primers designed based on our sampled variation of *A. munita* and *P. californicus* did not yield any products similar to sequences outside of *Papaver* indicating that these sequences do not represent S-like alleles previously undetected in the genus. The finding of up to four homologous products from single individuals of the diploid *P. californicus* suggests paralogous gene duplication following divergence from other sampled taxa that may partially explain the monophyletic clade of sequences from that species. Interestingly, each member of co-segregating PCR product pairs *S2d* and *S2e* as well as *S3a* and *S3b* are found in separate *P. californicus* sub-clades (Figure 2), suggesting that *S2d* and *S3a* may be paralogs to *S2e* and *S3b* respectively.

Gene duplication of S-like products is apparently common in the SI systems for which we now have molecular data. In the sporophytic system of the Brassicaceae the homologous polymorphic stigmatic S-locus genes *SRK* and *SLG* (S-locus glycoprotein) are both found in *Brassica oleracea*, *B. campestris* (Kusaba *et al*. 1997; Miege *et al*. 2001) and *Raphanus sativas* (Sakamoto *et al*. 1998) while only *SRK* is found in *Arabidopsis lyrata*, *A. halleri* (Schierup *et al*. 2001*a*; Castric and Vekemans 2004; Becshgaard *et al*. 2006) and *Capsella grandiflora* (Paetsch *et al*. 2006; Guo *et al*. 2009; Foxe *et al*. 2009). This implies that SLG polymorphism was either lost in some ancestor of both *Arabidopsis* and *Capsella* or gained in *Brassica/Raphanus*. Recently, Busch and Schoen (2008) also reported a co-segregating homologous polymorphism (‘*Lal*2’) linked to SRK in *Leavenworthia alabamica* whose function in SI is uncertain but is highly divergent from *SRK*. Using the SRK-SLG system as a model, Takuno *et al.* (2008) demonstrated that when a functional SI gene is under diversifying selection, a duplicated copy is advantageous because it can contribute to variation in the functional copy through gene conversion. This prediction was supported regardless of whether the second copy became a pseudogene and had no direct function in SI, especially if repeated exchanges with functional alleles has occurred. Because the paralogs found in *P. californicus* appear to represent complete open reading frames, we don't know if one or both copies are functional.

Paralogous duplications also appear abundant in the S-RNase system where multiple pollen-S-like F-Box genes are found linked to the S-locus in some taxa (Sassa *et al.* 2007; Newbigin *et al.* 2008). Presence of multiple linked loci makes it difficult to isolate products directly involved in SI (Wilson and Newbigin 2007). Theory predicts that pollen and stigma genes should co-evolve within haplotypes to avoid the breakdown of SI. However, only in some cases does it appear that divergence and history of pollen and pistil expressed recognition proteins are equivalent (Sato *et al.* 2002) while several others show remarkably dissimilar levels of divergence (Newbigin *et al.* 2008; Kohn 2008) between pollen and pistil genes, with low levels of divergence associated with duplications. This suggests that the type of gene conversion described by Takuno *et al.* (2008) may play a role in shaping genealogies and perhaps also in generating new specificities.

The lack of resolution in the phylogeny among *A. munita* sequences and the monophyletic clustering of those from *P. californicus* may reflect intra- and inter-locus recombination. The expected consequence of recombination on genealogies of balanced polymorphisms is shorter overall times to coalescence and long terminal branches relative to depth resulting in a topology with a star-like appearance (Schierup and Hein 2000; Shierup *et al*. 2001*b*). Although recombination is generally thought to be suppressed at the S-locus to preserve pollen-stigma gene interactions, studies that estimate recombination among S-alleles suggest that it may occur (Awadalla and Charlesworth 1999; Wang *et al*. 2001; Posada 2002; Vieira *et al.* 2003; Kamau *et al*. 2007).

If recombination has occurred where long tracts of flanking regions among homologous genes are exchanged, we might expect to see a correlation between LD and distance among a sample of alleles. Alternatively, if gene conversion has occurred where short tracts have been inserted within coding regions (Wiuf and Hein 2000; Takuno *et al*. 2008), tests for LD and distance may not be informative because LD estimates are obtained from sites that are near one another. The permutation test for recombination indicated that *A. munita* does not show significant correlations between LD and distance, but this may be due to the demonstrated low power of the test given the observed levels of sequence diversity in this species. The same test for *P. californicus* does indicate the presence of recombination, which may reflect exchanges between paralogous ends. Interestingly, estimates of the population recombination rate for *A. munita* are significantly greater than *P. californicus* and the estimate of the ratio of rate of recombination to mutation is higher. A higher recombination rate may explain the low resolution of internal nodes for sequences from *A. munita* relative to *P. californicus* while gene duplication and gene conversion in these taxa may explain both the low levels of variation and the monophyletic or near-monophyletic clusters of sequences observed in these genera relative to *Papaver*.

Population bottlenecks resulting from colonization or other demographic events have also been shown to affect levels of shared ancestral polymorphism and allelic variation of balanced polymorphisms. Richman *et al*. (1996) estimated ancient historic effective population in *Physalis crassifolia* that was an order of magnitude lower than *Solanum carolinense* based on reduced S-RNase lineages as a result of a putative bottleneck event. More recently Hasselman *et al*. (2008) found complete turnover (no shared ancestral polymorphism) in complementary sex determining alleles of three species of *Apis.* Lack of ancient polymorphism was ascribed to small long-term effective population sizes of bees. Miller *et al*. (2008) discovered greatly reduced trans-generic lineages among Old World lineages of *Lycium* S-RNases whose ancestors colonized Africa after long-distance dispersal from the New World. The patterns exhibited by *P. mcconnellii* putative S-alleles may be compatible with a colonization event to Alaska from Eastern Russia, though historical biogeographical data is lacking.

In general both PAML and OmegaMap selection estimates suggest that there is either a) substantial selective constraint and few residues are undergoing diversifying selection or b) the models for estimating selection are not able to detect important functional codon substitutions necessary for recognition (Hughes 2008; Yokoyama *et al*. 2008). The secondary structure of the *Papaver* S-alleles *PrS1, PrS3, PrS8* and *PnudSn1* proteins were determined by Kakeda *et al*. (1998) and predicted seven hydrophilic surface loops. Substitution at a single variable site within loop 6 of

PrS1 caused loss of function. This amino acid corresponds to site 107 in our alignment (Figure 1) and is predicted to be positively selected in all permutations of our datasets. Supporting functional assays as such refute criticisms made by Hughes (2007) that the types of statistical methods used here are unreliable in detecting important functional regions that are undergoing diversifying selection.

The selection models in PAML are considered ‘random effects likelihood’ (REL) models and use cumulative likelihoods of pre-defined rate classes to conclude whether there is selection on a proportion of sites in the alignment. The ability to pool data across sites increases the power to detect selection where site-by-site methods would not when alignments have few sequences or low divergence (Kosakovsky Pond *et al.* 2009), but this does not necessarily increase the ability to infer individual positively selected sites. Lastly, the potential effects of recombination and gene conversion cannot be ignored when interpreting our selection results. As Schierup, Mikkelson and Hein (2001) point out, when recombination has occurred, balancing selection is acting on shorter regions of recombinant sequences. Simulations conducted by Wilson and McVean (2006) showed that *codeml* in PAML gave false positives for sites under selection when recombination was detected in the dataset. This does not appear to be the case for our dataset as all sites identified as under selection using OmegaMap were also predicted using PAML.

We have presented sequences homologous to the *Papaver* S-locus from several naturally occurring species of Papaveraceae. While it is not yet proven that any of these sequences function in self-pollen recognition and rejection, the correspondence between genotype and incompatibility phenotype, and the molecular analysis of selection both suggest strongly that they may, or at the very least that some or all are linked to the S-locus in these taxa. The surprising phylogenetic relationships, with much lower levels of sequence displayed by *Argemone munita* and *Platystemon californicus* relative to the known levels of polymorphism among confirmed S-alleles from *Papaver* are somewhat surprising, though not completely unprecedented given the low levels of polymorphism seen at the pollen determining loci of some taxa with RNase-based incompatibility. If observed levels of polymorphism represent real differences at the S-locus among the genera reported here, then molecular details of the S-locus and the genome in which it resides, for instance whether or not there are paralagous copies of S-locus genes, may play a very large role in shaping patterns of polymorphism observed at loci under balancing selection.

Acknowledgements

Hastings Natural History Reserve, Sierra Nevada Aquatic Research Lab and UC Reserve system Gary Hannan, VE Franklin-Tong and Michael Lawrence, Sergei Kosakovsky Pond, Danny Wilson, California Desert Research Fund, Mildred Mathias Grant, Jeanne Messier Memorial Fund, NSF.

Table 1. Results of hand pollinations among *A. munita* full sibs with matching genotype (a). Genotype predicted the SI phenotype unless parents possessed allele *S25b* (bold). Parents of full sib Family 8-1 had genotype *S8a S8b* x *S1c S1d*. Full sib Family 25-4 parental genotypes were *S25a S25b x S4a S4d.* Predicted 'semi-compatible' hand pollinations (b) from both species where parents shared one common sequence (bold) set fruit 100% of the time.

Argemone munita

a) Family 8-1		Family 25-4		b) 'Semi-compatible'	
Genotype	**Fruit Set**	**Genotype**	**Fruit Set**	**Genotype**	**Fruit Set**
S8a S1c	0/7	*S4a S25a*	0/32	*S8a* ***S1c*** x *S8b* ***S1c***	7//7
S8b S1d	0/5	*S4a* ***S25b***	**8/16**	*S8a* ***S*1d** x *S8b* ***S1d***	7//7
*S8*b *S1c*	0/3	*S4d S25a*	0/11	*S4a* ***S25a*** x *S4d* ***S25a***	20/20
		S4d ***S25b***	**12/18**	***S*4a** *S25b* x ***S4a*** *S25b*	12//12
				S4a ***S25b*** x *S4*d ***S25b***	18/18
total	0/15		20/77	*total*	64/64

Table 2. Results of hand pollinations of *P. californicus* full sibs with matching genotypes (a). Parents of full sib Family 34-26 were genotyped *S1a S1b* x *S3a,b S3c_*where 3*a,b* appear to be a linked polymorphism in F1's and 3*c* is allelic to one or both *a,b* sequences (see text). Family 34-21 parents were of genotypes *S1a S1b* (as in previous family) x *S2d,e S2?*. Sequences 2*d,e* were always either inherited together in F1's or, when neither was present, no alternative product was found. Asterisks indicate that individuals may be homozygous for *S1b* or an alternative product was present but not found. Genotype *S1a S1b* was derived from non-matching allele individuals from F1's of 34-21 progeny. b) Predicted 'semi-compatible' hand pollinated crosses in which sibling parents shared one common sequence (bold). In total, 12% frutt set was observed following matching genotype crosses while 93% fruit set was observed when parents shared only one putative allele.

Platystemon californicus

a) Family 34-26		Family 34-21		b) 'Semi-compatible'	
Genotype	**Fruit Set**	**Genotype**	**Fruit Set**	**Genotype**	**Fruit Set**
S1a S3a,b	0/33	*S1a S2d,e*	4/18	***S1a*** *S3a,b* x ***S1a*** S3*c*	16/17
S1b S3a,b	0/17	*S1b S2d,e*	1/13	***S1b*** *S3ab* x ***S1b*** S3*c*	18/20
*S1a S3c_*	1/8	*S1b S2?*	--	S1*a* ***S3a,b*** x ***S1b*** S3*a,b*	7//7
S1b S3?	10/23	F2's		***S1b*** S2*d,e* x ***S1b*** S2?	5/5
		S1a S1b	0/17	S1*a* ***S2d,e*** x S1*b* ***S2d,e***	10//11
total	11/81		5/48	*total*	56/60

Table 3. Statistics of combined full sib crosses from each species. Matching genotype indicates both parents possess the same alleles while Semi indicates only one sequence was shared among parents in the cross. The null hypothesis that genotype does not predict fruit set is strongly rejected for each species ($p << 0.001$) by $\chi2$.

A. munita	**Fruit Set**			*P. californicus*	**Fruit Set**		
Genotype	**Yes**	**No**	**# Crosses**	**Genotype**	**Yes**	**No**	**# Crosses**
Matching	20	72	92	Matching	16	113	129
Semi	64	0	64	Semi	56	4	60
Total	84	72	156	Total	72	117	189
	$\chi2 = 93.02$ (1 *df*) $p < 0.001$				$\chi2 = 113.74$ (1 *df*) $p < 0.001$		

Table 4. Average pairwise nucleotide divergence among sequences within genera. For all taxa only sequences with 5' end were included except *R. coulteri* which has only 3' end sequences. The four known *Papaver* S-alleles have the greatest overall divergence for all types of sites.

	Synonymous (Π_s)	**Non-Synonymous (Π_a)**	**All Sites (Π)**
Papaver (S1, S3, S8, Sn1)	0.7	0.28	0.36
Papaver (all)	0.57	0.24	0.3
A. munita	0.3	0.14	0.17
P. californicus	0.49	0.2	0.26
R. coulteri	0.2	0.11	0.13

Table 5. Results of PAML analysis of positive selection on 5’ sequences from *Papaver*, *Argemone* and *Platystemon*. a) Sequence data from *Papaver* included. Models allowing for selection provided significantly better fits than neutral models. except the general discrete M3 model. b) *Papaver* sequences excluded.

a) 5’ Alignment of *Papaver* (10), *A. munita* (11) and *P.californicus* (10) sequences

Model	**Likelihood**	**Sites (posterior score)**	**Mean** w, **Standard Error**
M1 (neutral)	-6410.82	N/A	
M2 (selection)	-6406.15	**107 (0.946)**	**2.430 +- 0.456**
		119 (0.974)	2.470 +- 0.394
M3 (discrete)	-6374.01	No sites	
M7 (beta dist.)	-6370.63	N/A	
M8 (beta and w)	-6366.94	53 (0.8)	1.404 +- 0.311
		107 (0.92)	1.490 +- 0.270
		119 (0.95)	1.514 +- 0.247

b) 5’Alignment of *A. munita* (11), *P.californicus* (10) sequences only (No *Papaver*)

Model	**Likelihood**	**Sites (posterior score)**	**Mean** W, **Standard Error**
M1 (neutral)	-4055.71	N/A	
M2 (selection)	-4049.04	53 (0.99)	2.779 +- 0.545
		107(0.911)	2.638 +- 0.710
		119 (0.97)	2.729 +- 0.610
M3 (discrete)	-4034.8	53(1.00)	1.782
		77 (0.98)	1.778
		83 (0.97)	1.745
		103 (0.99)	1.765
		107 (0.997)	**1.778**
		119 (0.998)	1.78
M7 (beta dist.)	-4037.95	N/A	
M8 (beta and w)	-4030.48	53 (0.99)	2.204 +- 0.499
		107 (0.93)	**2.132 +- 0.574**
		119 (0.958)	2.167 +- 0.540

Table 6. Results of non-parametric permutation analyses of recombination and coalescent likelihood estimates of $\rho = 4N_er$. The null hypothesis of no recombination was rejected using all three linkage disequilibrium statistics for *P. californicus* but not *A. munita*. The overall population recombination rate ρ is significantly greater for *A. munita*. Mean estimates of the population mutation rate θ and ρ are reported with 95% highest posterior density (HPD) distributions. We then estimated the ratio of recombination rate *r* and μ from mean ρ/ θ.

	Argemone munita		***Platystemon californicus***	
	Correlation	*P*-value	Correlation	*P*-value
r^2	-0.0147	0.129	-0.0689	0.001
D'	-0.0261	0.139	-0.0533	0.014
G4	-0.0212	0.202	-0.0426	0.035
	Mean (95% HPD)		Mean (95% HPD)	
θ	0.94 (0.78, 1.15)		2.27 (1.93, 2.69)	
ρ	3.01 (2.49, 3.61)		1.54 (1.23, 1.94)	
r/μ	3.2		0.68	

Table 7. Simulation results testing the power to detect recombination under variable levels of genetic diversity (θ) with a sample size of 24 sequences. The power increases as θ approaches the value determined for *P. californicus* estimates (Table 7).

θ	Frequency of rejecting null H_0: $4N_er = 0$
0.5	0.429
1	0.651
1.5	0.8
2	0.9
2.5	0.914
3	0.946

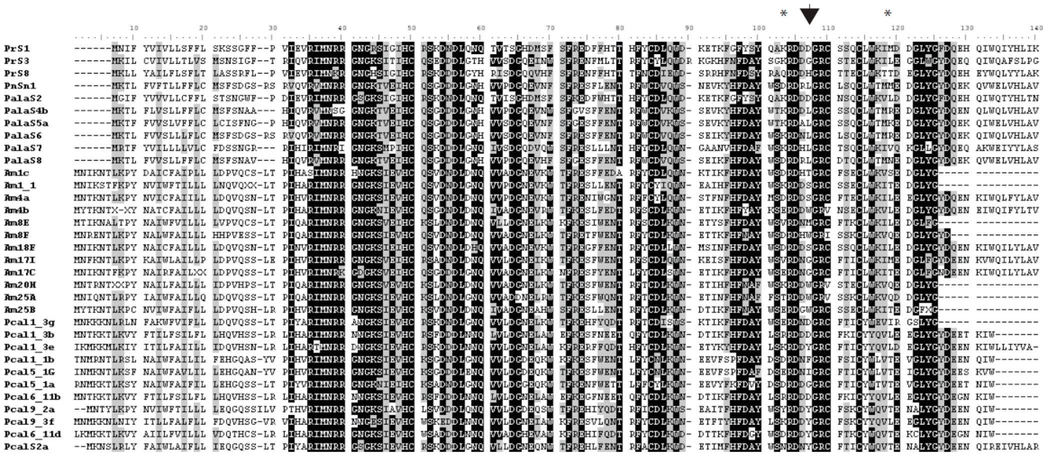

Figure 1. Amino acid alignment of all sequences containing 5' ends. All *Papaver* spp. alleles (10) shown are full length possessing both 5'and 3' ends. Asterisks indicate sites estimated to be under positive selection in both PAML (YANG 2000) and OmegaMap (WILSON and MCVEAN 2006) analyses. Black arrow indicates the amino acid (site 107) used in the site directed mutagenesis experiment of KAKEDA *et al.* (1998; see text) and is also predicted to be under positive selection.

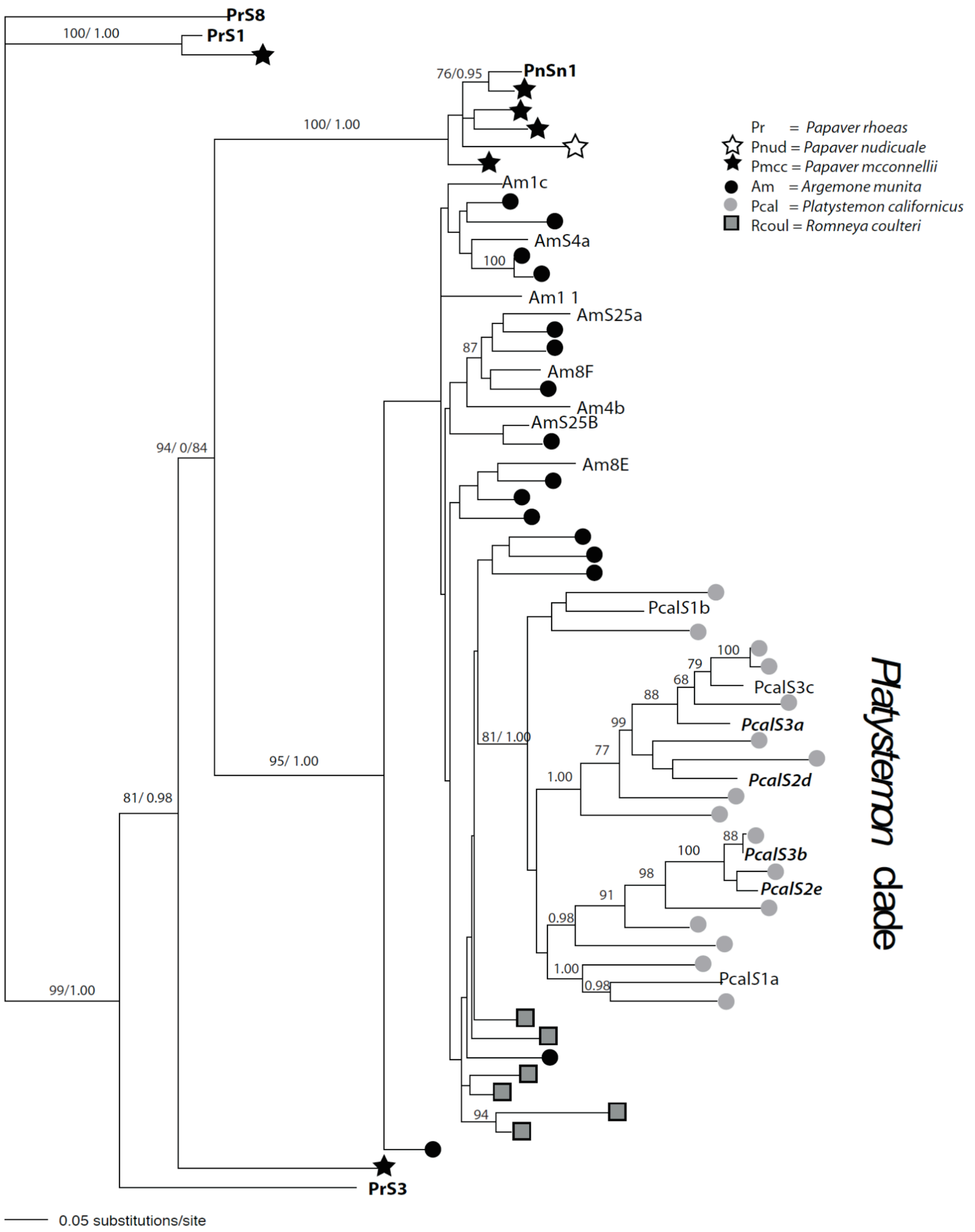

Figure 2. Maximum likelihood phylogeny of known *Papaver* S-alleles and putative S-alleles from Alaskan and Californian taxa. The *P.rhoeas* S-alleles (PrS1, PrS3 and PrS8 in bold) identified by Foote et al (1994) form highly divergent lineages that include recently identified sequences from *P. mcconnellii* no close associations with any sequences from outside of the genus. All sequences from *P. californicus* form a distinct monophyletic clade while sequences from *A. munita*, *R. coulteri* show some shared polymorphism.

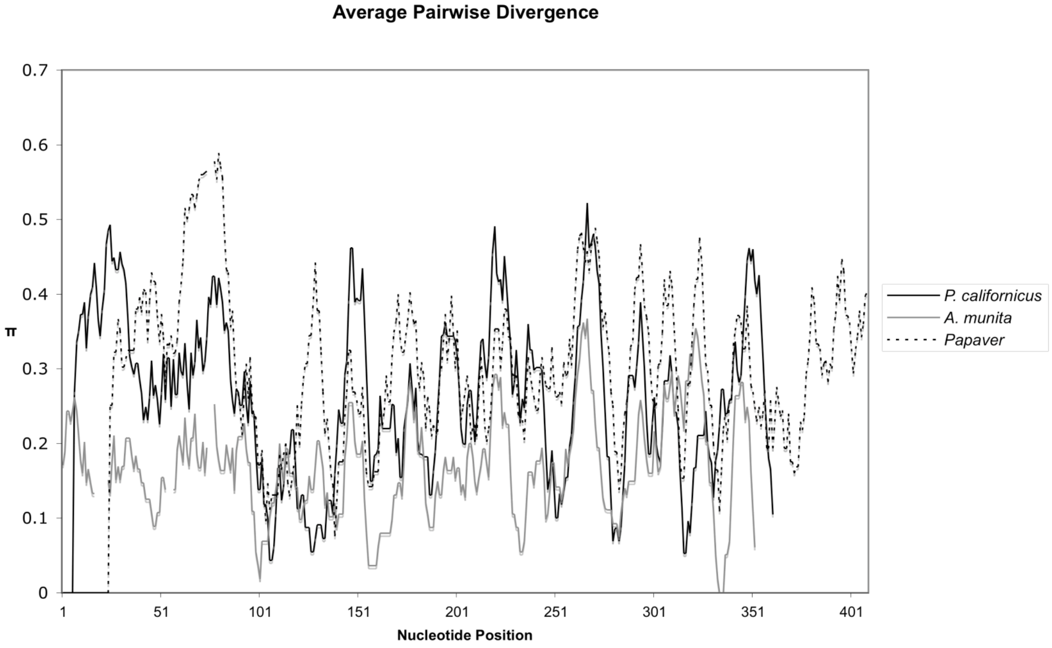

Figure 3. Sliding window analysis of average pairwise nucleotide diversity (π) of putative S-alleles. Data consist of 10 *Papaver*, 10 *P. californicus,* and 11 *A. munita* sequences. Only sequences possessing 5' end were used. (Check Non-synonymous and synonymous sliding windows as well for checking PAML). Divergence was estimated using DNASP (Rozas *et al*. 2003).

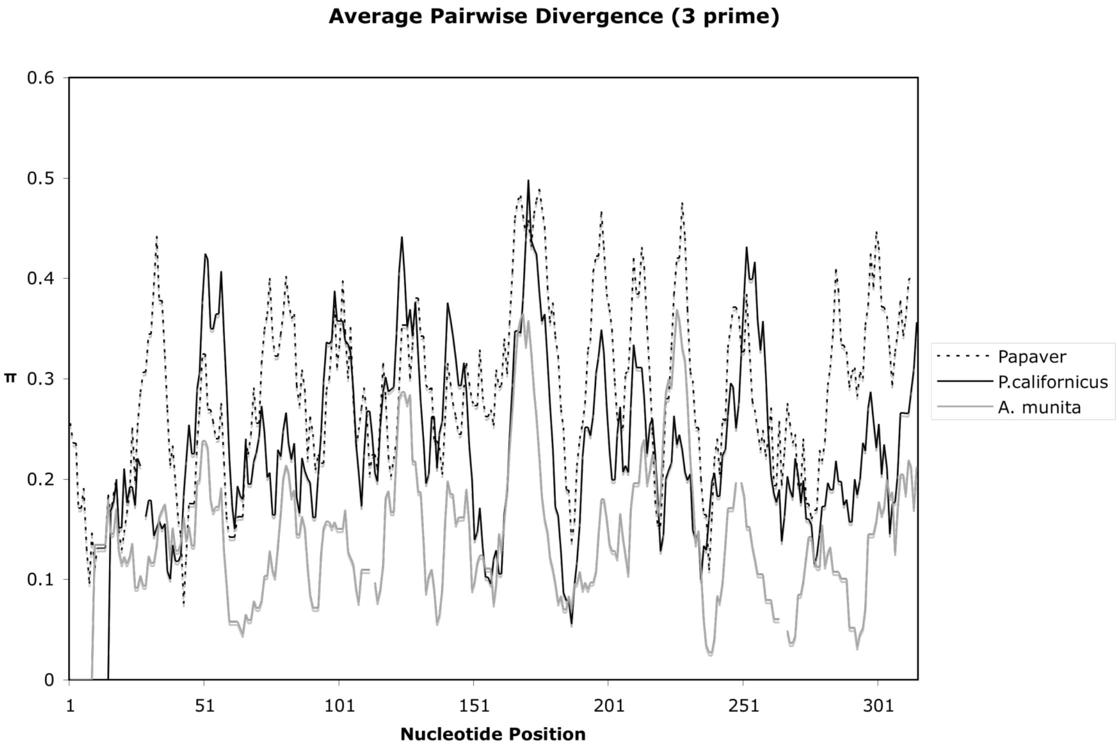

Figure 4. Sliding window analysis of average pairwise nucleotide diversity (π) of putative S-alleles. Data comprise 10 *Papaver*, 24 *P. californicus,* and 28 *A. munita* sequences. Only sequences possessing complete 3' end were used. Divergence was estimated using DNASP (Rozas *et al*. 2003).

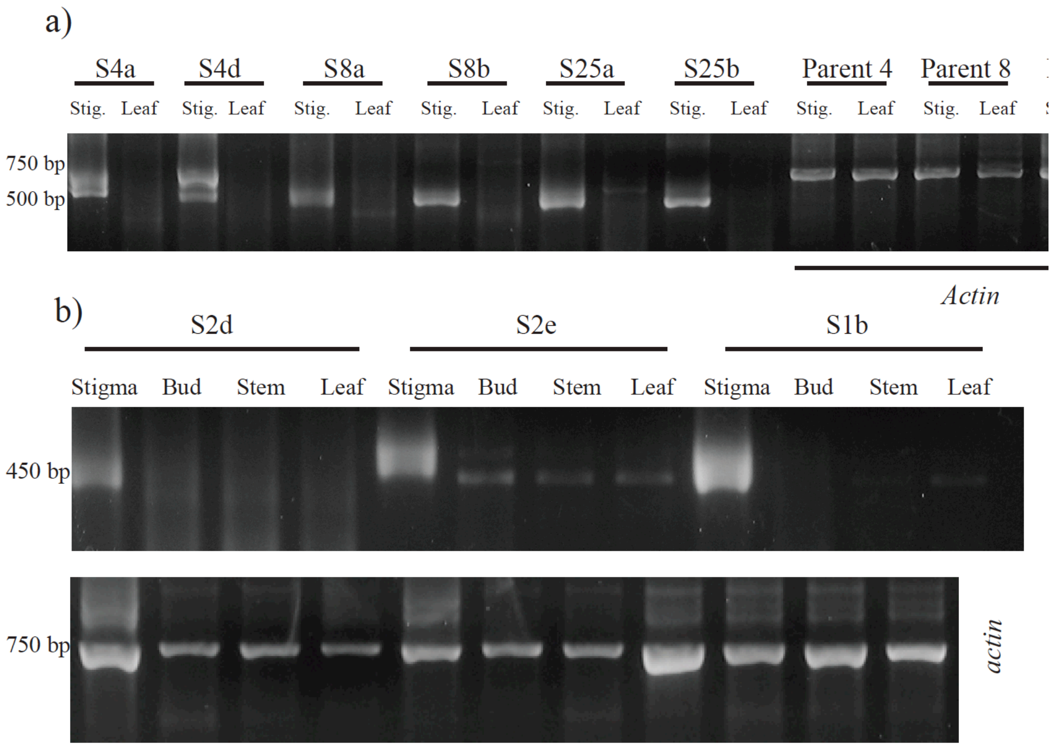

Figure 5. RT-PCR of putative S-locus genes from stigmatic, stem and leaf tissues using allele-specific forward primers and a universal amplification reverse primer that binds to 3' end of the mRNA poly-A tail. Expression of any allele was only detected in stigmatic tissue of *A. munita* (a) and not in leaf tissue. *Actin* primers for all parent tissues were used as positive controls. Expression in *P. californicus* tissues is highest in stigmas as expected for S-alleles (b). Putative alleles *S2d* and *S2e* are paralogs. Sequencing of bands in non-stigmatic tissue using *S2e* primers showed that they are products not previously identified from stigmas.

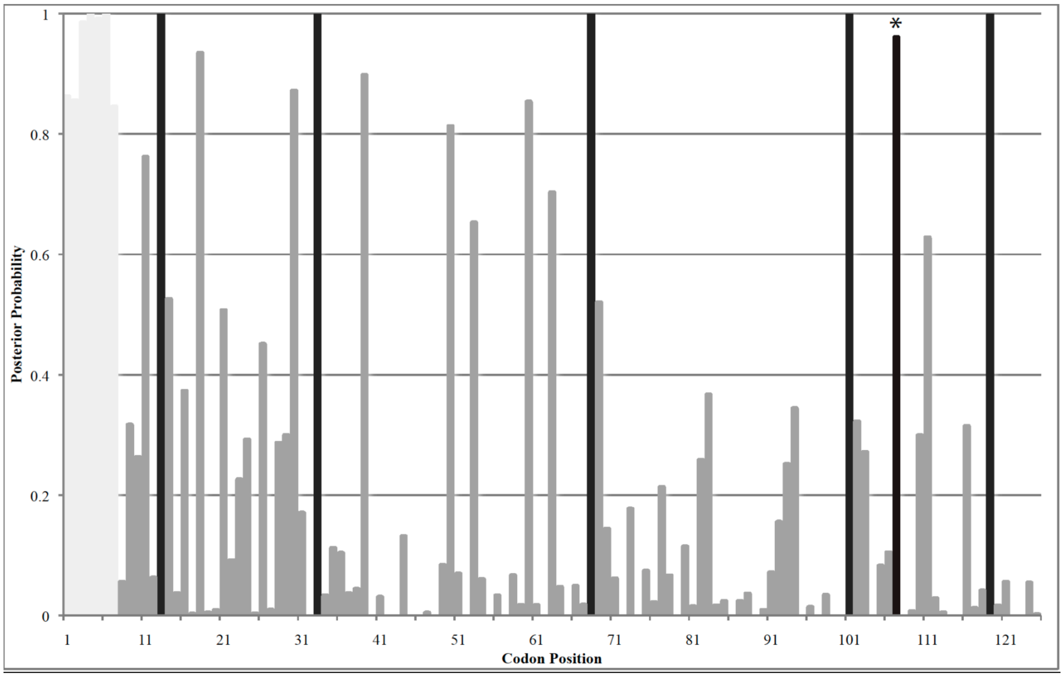

Figure 6. Posterior probabilities for codons estmatied to be under positive selection using OmegaMap (Wilson and McVean 2006) implemented with an independent model of rate variation for each site. The dataset used includes 10 *Papaver*, 10 *P. californicus* and 12 *A. munita* sequences. Sites 1-7 from *A. munita* and *P. californicus* are excluded because they are not present in *Papaver*. Codons 13, 33, 68, 103, 119 (black bars) and 107 (∗) are predicted to be under positive selection.

References

Awadalla P, Charlesworth D.1999. Recombination and selection at *Brassica* self-incompatibility loci. Genetics 152: 413–425.

Beatty, A. V. 1936. Genetic studies on the California Poppy. J. Heredity 27:330-338.

Bechsgaard, J.S., V. Castric, D. Charlesworth,_ X.Vekemans, and M.H. Schierup. 2006. The Transition to self-compatibility in *Arabidopsis thaliana* and evolution within S-Haplotypes over 10 Myr. Mol. Biol. Evol. 23:1741–1750.

Busch, JW., J. Sharma and D. J. Schoen. 2008. Molecular Characterization of Lal2, an SRK-Like Gene Linked to the S-Locus in the Wild Mustard *Leavenworthia alabamica*. Genetics 178: 2055–2067

Bosch, M. and VE. Franklin-Tong. 2008. Self-incompatibility in *Papaver*: signalling to trigger PCD in incompatible pollen. Journal of Experimental Botany. 59: 481–490

Castric, V. and X. Vekemens. 2004. Plant self-incompatibility in natural populations: a critical assessment of recent theoretical and empirical advances. Molecular Ecology. 13: 2873–2889

Cook, S. 1962. Genetic system, variation and adaptation in *Eschscholzia californica*. Evolution. 16: 278-299

Charlesworth, D, BK. Mable, MH. Schierup, C. Bartolome and P. Awadalla. 2003. Diversity and Linkage of Genes in the Self-incompatibility gene family in *Arabidopsis lyrata*. Genetics. 164: 1519–1535

de Graaf, B. H. J., JJ Rudd, MJ Wheeler, RM. Perry, EM. Bell, K. Osman, F. C. H. Franklin1 & V E Franklin-Tong. 2006. Self-incompatibility in *Papaver* targets soluble inorganic pyrophosphatases in pollen. 444: 490-493

deNettancourt D. 1977. Incompatibility in Angiosperms. New York: Springer Press

Doyle, J.J. and Doyle, J.L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissues. Phytochemical Bulletin 19: 11-15.

Foote, H. G., Ride, J. P., Franklin-Tong, V. E., Walker, E. A., Lawerence, M. J. & Franklin, F. C. H. 1994 Cloning and expression of a distinctive class of self-incompatibility (*S*-) gene from *Papaver rhoeas*. PNAS. 91: 2265-2269

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. 2009. Recent speciation associated with the evolution of selfing in *Capsella*. Proc Nat. Acad. Sci Feburary

Franklin-Tong, V.E. and F. Chris H. Franklin. 2003. Gametophytic self-incompatibility inhibits pollen tube growth using different mechanisms. Trends in Plant Sciences. 8: 598-605

Guo, Y, J.S. Bechsgaard, T.Slotte, B.Neuffer, M. Lascoux, D. Weigel, and M. H. Schierup. 2009. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck Proc. Nat. Acad. Sci.106: 5246-5251

Hannan, G. 1981. Flower color polymorphism and pollination biology of *Platystemon californicus*. American Journal of Botany. 68: 233-243

Hasselmann, M., X.Vekemans, J.Pflugfelder, N. Koeniger, G. Koeniger, S. Tingek, and Martin Beye. 2008. Evidence for Convergent Nucleotide Evolution and High Allelic Turnover Ratesat the complementary sex determiner Gene of Western and Asian Honeybees. Mol. Biol. Evol. 25: 696–708.

Hauck, NR. K. Ikeda, R. tao andAF Iezzoni 2006. The mutated S1-haplotype in sour cherry has an sltered S-haplotype–specific F-Box protein gene. Heredity 2006:97: 514–520

Hoot, SB, JW Kaderiet, FR Blattner, KB Jork, AE Schwarzbach, PR Crane. 1997. Data congruence and phylogeny of the Papaveraceae s.l. Based on four data sets: atpB and rbcL sequences, trnK restriction sites and morphological characters. Systematic Botany. 22: 575-590

Hudson, R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17: 754–755.

Hughes, A. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. Heredity. 99: 364-373.

Hughes, A. 2008. The origin of adaptive phenotypes. Proc. Nat. Acad. Sci. 105:13193-13194

Igic, B., L.Bohs and JR Kohn. 2004. Historical inferences from the self-incompatibility locus. New Phytologist 161: 97–105

Igic B., W.A. Smith, K. Robertson, B.A. Schaal, and J.R. Kohn. 2007. The population genetics of the self-incompatibility polymorphism in wild tomatoes: I. S-RNase diversity in *Solanum chilense* (Dun.) Reiche (Solanaceae). Heredity 99:553-561

Ikeda K, Igic B, Ushijima K, Yamane H, Hauck NR, Nakano R, Sassa H, Iezzoni AF, Kohn JR, Tao R. 2004. Primary structural features of the S haplotype-specific F-box protein, SFB, in *Prunus*. Sex Plant Reprod 16:235–243

Ioerger, TR, AG Clark, T-h Kao. 1990. Polymorphism at the Self-Incompatibility Locus in Solanaceae Predates Speciation PNAS 1990; 87: 9732-9735.

Kakeda K, Jordan ND, Conner A, Ride JP, Franklin-Tong VE, Franklin FC. 1998. Identification of residues in a hydrophilic loop of the Papaver rhoeas S protein that play a crucial role in recognition of incompatible pollen. Plant Cell. 10:1723-32

Kamau, E., B.Charlesworth and D. Charlesworth. 2007. Linkage disequilibrium and recombination rate estimates in the self-incompatibility region of Arabidopsis lyrata. Genetics 176: 2357–2369

Kohn, JR. 2008. What genealogies of S-alleles tell us. in Self-incompatibility in Flowering Plants: Evolution, diversity and mechanisms. Springer-Verlag Berlin Heidelberg. pp103-117

Kosakovsky Pond, S.L, S. D. W. Frost and S.V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676-679

Kosakovsky Pond, S.L Art F.Y. Poon, and Simon D.W. Frost. 2009. Estimating selection pressures on alignments of coding sequences: Analyses using *HyPhy*. The Phylogenetic Handbook. Cambridge University Press. in press

Kurup S, Ride JP, Jordan N, Fletcher G, Franklin-Tong VE, Franklin FCH. 1998. Identification and cloning of related self-incompatibility S-genes in *Papaver rhoeas* and *Papaver nudicaule*. Sexual Plant Reproduction 11, 192±198.

Kusaba, M., T. Nishio, Y. Satta, K. Hinata, and D. Ockendon. 1997. Striking sequence similarity in inter- and intra-specific comparisons of class I *SLG* alleles from *Brassica oleracea* and *Brassica campestris*: Implications for the evolution and recognition mechanism. PNAS. 94: 7673-7678

Kusaba, M., K. Dwyer, J. Hendershot, J. Vrebalov, J. B. Nasrallah *et al.*, 2001. Self-incompatibility in the genus Arabidopsis: characterization of the S locus in the outcrossing A. lyrata and its autogamous relative A. thaliana. Plant Cell 13: 627–643.

Lane MD and Lawrence MJ. 1993. The population genetics of the self-incompatibility
polymorphism in *Papaver rhoeas.* VII. The number of S-alleles in the species. Heredity 71: 596-602.

Lawrence, MJ and S. O’Donnell. 1981. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. III. The number and frequency of S-alleles in two further natural populations. Heredity. 47: 53-61

Mable, B. K., M. H. Schierup, and D. Charlesworth. 2003. Estimating the number, frequency, and dominance of S-alleles in a natural population of *Arabidopsis lyrata* (Brassicaceae) with sporophytic control of self-incompatibility. Heredity 90: 422–431.

Mable, BK. 2004. Polyploidy and self-compatibility: is there an association? New Phytologist. 163: 803-811.

Meunier, J. and A. Eyre-Walker. 2001 The correlation between linkage disequilibrium and distance. Implications for recombination in Hominid mitochondria. Mol. Biol. Evol. 18:2132-213

Miller JS, Levin RA, and Feliciano NM. 2008. A tale of two continents: Baker’s rule and the maintenance of selfincompatibility in Lycium (Solanaceae). Evolution. 62-5: 1052–1065

McVean, G. A. T., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241

Miege, C., VR Chaˆble, MH Schierup, and D Cabrillac. Intra-haplotype Polymorphism at the *Brassica S* Locus. 2001. Genetics. 159: 811–822

Newbigin, E., T. Paape**,** and J.R. Kohn. 2008. RNase based self-incompatibility: Puzzled by Pollen S. Plant Cell. 20:2286-2292

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–36.

O'Donnell S, Lawrence MJ. 1993. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas.* VI. Estimation of the overlap between allelic complements of a pair of populations. Heredity. 71: 591-595

Paape, T**.,** B. Igic, S. Smith, R. Olmstead, L. Bohs, J.R. Kohn. 2008. A 15-Million-Year-Old Genetic Bottleneck at the S-locus of the Solanaceae. Mol. Biol. Evol. 25: 655-663

Paetsch, M., S. Mayland-Quellhorst and B. Nueffer. 2006. Evolution of the self-incompatibility system in the Brassicaceae: identification of S-locus receptor kinase (SRK) in self-incompatible *Capsella grandiflora*. Heredity 97: 283–290

Posada D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. 19:708–17

Posada, D., and K. A. Crandall. 2002. The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. 54:396–402

Qiao, H., F. Wang, L. Zhao, J. Zhou, Z. Lai, Y. Zhang, TP Robbins, and Y. Xuea. 2004. The F-Box Protein AhSLF-S2 Controls the Pollen Function of S-RNase–Based Self-Incompatibility. The Plant Cell. 16: 2307–2322

Rambaut, A. and Grassly, N. C. 1996. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci.

Richman, AD, MK Uyenoyama and JR. Kohn. 1996. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science. 273: 1212-1216

Richman AD and Kohn JR. 2000. Evolutionary genetics of self-incompatibility in the Solanaceae. Plant Mol Biol 42: 169–179

Ride, JP., EM Davies, FCH Franklin, DF Marshall. 1999. Analysis of *Arabidopsis* genome sequence reveals a large new gene family in plants. Plant Molecular Biology. 39**:** 927–932

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Rudd JJ, and VE Franklin-Tong. 2003. Signals and targets of the self-incompatibility response in pollen of *Papaver rhoeas*. Journal of Experimental Botany 54: 141-148. Plant Reproductive Biology Special Issue

Sakamoto K, Kusaba M, Nishio T. 1998. Polymorphism of the S-locus glycoprotein gene (SLG) and the S-locus related gene (SLR1) in *Raphanus sativus* L. and self-incompatible ornamental plants in the Brassicaceae. Mol Gen Genet 258: 397–403

Sassa, H., Kakui, H., Miyamoto, M., Suzuki, Y., Hanada, T., Ushijima, K., Kusaba,., Hirano, H., and Koba, T. 2007. S locus F-box brothers: Multiple and pollen-specific F-box genes with S haplotype-specific polymorphisms in apple and Japanese pear. Genetics 175, 1869-1881.

Sato, T., R. Fujimoto, K. Toriyama and T. Nishio, 2003 Commonality of self-recognition specificity of S haplotypes between *Brassica oleracea* and *Brassica rapa.* Plant Mol. Biol. 52**:** 617–626.

Savage, A. E., and J. S. Miller. 2006. Gametophytic self-incompatibility in *Lycium parishii* (Solanaceae): allelic diversity, genealogical structure, and patterns of molecular evolution at the S-RNase locus. Heredity 96:434–444.

Schierup, M. H., and J. Hein. 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879–891

Schierup, M. H., B. K. Mable, P. Awadalla and D. Charlesworth. 2001a. Identification and characterization of a polymorphic receptor kinase gene linked to the self-incompatibility locus of *Arabidopsis lyrata*. Genetics 158: 387–399

Schierup, M. H., A. M. Mikkelsen, and J. Hein. 2001*b*. Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genetics 159:1833–1844

Sijacic, P., Xi Wang, Andrea L. Skirpan, Yan Wang, Peter E. Dowd, Andrew G. McCubbin, Shihshieh Huang and Teh-hui Kao. 2004. Identification of the pollen determinant of S-RNase-mediated self-incompatibility. Nature. 429: 302-305

Solstad, H. 2008. Taxonomy and evolution of the diploid and polyploid *Papaver* sect. *Meconella* (Papaveraceae). PhD Dissertation. University of Oslo.

Sonneveld, T, Robbins, TP, Bošković, R, Tobutt, KR. 2001. Cloning of six cherry self-incompatibility alleles and development of allele-specific PCR detection. Theoretical and Applied Genetics.102: 1046-1055

Stone, J. L., AND S. E. Pierce. 2005. Rapid recent radiation of S-RNase lineages in *Witheringia solanacea* (Solanaceae). Heredity 94: 547–555.

Swofford, DL. 2002. PAUP. Phylogenetic Analysis Using Parsimony (and other methods) Version 4. Sinauer: Sunderland, MA

Takebayashi, N., P. B. Brewer, E. Newbigin, and M. K. Uyenoyama. 2003. Patterns of variation within self-incompatibility loci. Mol. Biol. Evol. 20:1778-1794.

Takuno, S., T. Nishio, Y. Satta, and H. Innan. 2008. Preservation of a Pseudogene by Gene Conversion and Diversifying Selection. Genetics 180: 517-531

Testa G, Caccia R, Tilesi F, Soressi GP, Mazzucato A. 2002. Sequencing and characterization of tomato genes putatively involved in fruit set and early development. Sex Plant Reprod 14:269–277

Thomas, S., K. Osman, B.H.J. deGraaf, G. Shevchenko, M. Wheeler, F.C.H. Franklin, V.E. Franklin-Tong. 2003. Investigating mechanisms involved in the self-incompatibility response in *Papaver rhoeas*. Phil. Trans. R. Soc. Lond. B 358: 1033–1036

Thomas, SG, and VE Franklin-Tong. 2004. Self-incompatibility triggers programmed cell death in *Papaver* pollen. Nature. 429: 305-308

Turner LM, Hoekstra HE .2006. Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (*Peromyscus*). Molecular Biology and Evolution. 23:1656–1669.

Vieira CP and D Charlesworth. 2002. Molecular variation at the self-incompatibility locus in natural populations of the genera *Antirrhinum* and *Misopates.* Heredity 88: 172–181

Vieira CP, Charlesworth D and J Vieira. 2003. Evidence for rare recombination at the gametophytic self-incompatibility locus. Heredity 91:262-267.

Vieira, J., Morales-Hojas, R., Santos, R.A.M., and Vieira, C.P. 2007. Different positively selected sites at the gametophytic self-incompatibility pistil S-RNase gene in the Solanaceae and Rosaceae (*Prunus*, *Pyrus*, and *Malus*). J. Mol. Evol. 65, 175-185.

Walker EA, Ride JP, Kurup S, Franklin-Tong VE, Lawrence MJ, Franklin FCH. 1996. Molecular analysis of two functional homologues of the S3 allele of the *Papaver rhoeas self* incompatibility gene isolated from different populations. Plant Molecular Biology 30: 983-994

Wang, X., AL Hughes, T. Tsukamoto, T. Ando, and TH. Kao. 2001. Evidence That Intragenic Recombination Contributes to Allelic Diversity of the S-RNase Gene at the Self-Incompatibility (*S*) Locus in *Petunia inflata*. Plant Physiology. 125: 1012-1022

Wheeler, MJ, VE Franklin-Tong and VCH Franklin. 2001. The molecular and genetic basis of pollen-pistil interactions. New Phytologist. Tansley Review. 128: 565-580

Wilson DJ and McVean G. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics.172:1411–1425

Wiuf, C., and J. Hein, 2000. The coalescent with gene-conversion. Genetics 155: 451–462

Wheeler, D. and E. Newbigin. 2007. Expression of 10 S-Class SLF-like Genes in *Nicotiana alata* Pollen and Its Implications for Understanding the Pollen Factor of the S Locus. Genetics 177: 1–10

Wright, G. M. 1979. Self-incompatibility in *Eschscholzia californica*. Heredity 43 :429-431

Yamane, H., Ikeda, K., Ushijima, K., Sassa, H., and Tao, R. 2003. A pollen-expressed gene for a novel protein with an F-box motif that is very tightly linked to a gene for S-RNase in two species of cherry, Prunus cerasus and P. avium. Plant Cell Physiol. 44, 764–769.

Yang Z. 2000. Phylogenetic analysis by maximum likelihood (PAML). London: University College.

Yang Z, and Swanson WJ. 2002. Codon substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19:49-57

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–18.

Yokoyama, S., T. Tada, H. Zhang, and L. Britt. 2008. Elucidation of phenotypic adaptations: Molecular analysis of dim-light vision proteins in vertebrates. Proc. Nat. Acad. Sci. 105: 13480-13485