

Genomic Divergence during Speciation Driven by Adaptation to Altitude

Mark A. Chapman,^{‡,1} Simon J. Hiscock,² and Dmitry A. Filatov^{*,1}

¹Department of Plant Sciences, University of Oxford, Oxford, United Kingdom

²School of Biological Sciences, University of Bristol, Bristol, United Kingdom

[‡]Present address: Faculty of Natural and Environmental Sciences, University of Southampton, Highfield Campus, Southampton, United Kingdom

*Corresponding author: E-mail: dmitry.filatov@plants.ox.ac.uk

Associate editor: Stephen Wright

Abstract

Even though Darwin's "On the Origin of Species" implied selection being the main driver of species formation, the role of natural selection in speciation remains poorly understood. In particular, it remains unclear how selection at a few genes can lead to genomewide divergence and the formation of distinct species. We used a particularly attractive clear-cut case of recent plant ecological speciation to investigate the demography and genomic bases of species formation driven by adaptation to contrasting conditions. High-altitude *Senecio aethnensis* and low-altitude *S. chrysanthemifolius* live at the extremes of a mountain slope on Mt. Etna, Sicily, and form a hybrid zone at intermediate altitudes but remain morphologically distinct. Genetic differentiation of these species was analyzed at the DNA polymorphism and gene expression levels by high-throughput sequencing of transcriptomes from multiple individuals. Out of ~18,000 genes analyzed, only a small number (90) displayed differential expression between the two species. These genes showed significantly elevated species differentiation (F_{ST} and D_{xy}), consistent with diversifying selection acting on these genes. Genomewide genetic differentiation of the species is surprisingly low ($F_{ST} = 0.19$), while ~200 genes showed significantly higher (false discovery rate < 1%; mean outlier $F_{ST} > 0.6$) interspecific differentiation and evidence for local adaptation. Diversifying selection at only a handful of loci may be enough for the formation and maintenance of taxonomically well-defined species, despite ongoing gene flow. This provides an explanation of why many closely related species (in plants, in particular) remain phenotypically and ecologically distinct despite ongoing hybridization, a question that has long puzzled naturalists and geneticists alike.

Key words: adaptation, ecological speciation, hybrid zone, transcriptomics, demography.

Introduction

Understanding how species arise is a long-standing goal of evolutionary biology. Contrary to Darwin's view of the *Origin of Species by Means of Natural Selection* (Darwin 1859), speciation does not always occur due to natural selection, as species may arise via gradual accumulation of differences in isolated populations or via saltational change in chromosome number/structure causing immediate reproductive isolation between a pair of neospecies. In fact, this nonselectionist view of the speciation process dominated evolutionary biology over a significant proportion of last century (Mayr 1942). More recently, the evidence in favor of selection-based ecology-driven speciation has started to accumulate (Schluter 2009; Nosil 2012), and the focus of speciation studies has shifted to the analysis of closely related species that are still exchanging genes (Wolf et al. 2010; Smadja and Butlin 2011; Feder et al. 2012). Such intermediate stages of the speciation process have the potential to shed light on the very moment of speciation (as opposed to later stages of gradual divergence) and to reveal the relative importance of a range of factors in the development of reproductive isolation,

including locus-specific and genomewide divergent selection, and genetic drift (Schluter 2009; Nachman and Payseur 2012; Nosil 2012).

Ecological speciation gives rise to populations that are locally adapted such that if an individual from one population finds itself in the habitat of the other then it is less fit ("immigrant inviability"; Nosil et al. 2005). At the genetic level, the alleles for local adaptation are unfit in the alternate environment and therefore remain solely or mostly within the species in which they originated. Other loci, which do not affect ecological differences, are not subject to such strong diversifying selection and therefore, at least in the early stages of speciation, will not show such strong interspecific differentiation, creating a heterogeneous pattern of divergence throughout the genome and so-called speciation islands (Wu 2001; Nosil et al. 2009).

This mosaic of divergence between species has been uncovered in several species (e.g., Rieseberg et al. 1999; Scotti-Saintagne et al. 2004; Lawniczak et al. 2010; reviewed in Nosil et al. 2009) and may be particularly prominent if gene flow between the species is ongoing. In such situations, most notably hybrid zones, hybridization allows alleles at neutral

loci to introgress and recombine, whereas loci under divergent selection will be slowed or even prevented from introgressing (e.g., Martinsen et al. 2001; Payseur and Nachman 2005; Bull et al. 2006; Minder and Widmer 2008; Kane et al. 2009; Kulathinal et al. 2009; Stölting et al. 2013). Loci with patterns of differentiation that exceed the neutral expectation are candidates for underlying interspecific differences (Beaumont and Balding 2004; Storz 2005). Targets of selection may well be surrounded by an extended region of elevated differentiation, leading to the possibility that markers close to the target also show the signature of selection (Feder et al. 2012).

Altitudinal clines and hybrid zones present particularly convenient models to study adaptation and speciation because conditions and selective pressures change quite dramatically over fairly short distances (Payseur 2010). This study focuses on a plant hybrid zone on the slopes of Mt. Etna, Sicily, where two closely related species of ragwort (*Senecio*, Asteraceae), adapted to high- and low-altitude environments, hybridize at intermediate altitudes, but remain phenotypically distinct. High-altitude *Senecio aethnensis* Jan ex DC. is endemic to Mt. Etna, and is restricted to altitudes above ~2,000 m. Low-altitude *S. chrysanthemifolius* Poir. is found on the lower slopes of Mt. Etna (typically below ~1,000 m) and throughout Sicily, even reaching the southern tip of mainland Italy. Morphologically variable hybrids are found at intermediate altitudes on Mt. Etna (James and Abbott 2005; Brennan et al. 2009), and for traits that differentiate the parental species (most conspicuously, leaf dissection, inflorescence size, and flowering time), clines are found as one moves from one end of an altitudinal transect to the other (Crisp 1972). Analysis of putatively neutral molecular markers also demonstrates ongoing hybridization and backcrossing, with markers apparently able to introgress with relative ease between the species (James and Abbott 2005; Brennan et al. 2009). These findings are in line with the observation that the species are completely interfertile (Chapman et al. 2005; Brennan et al. 2013) and hybrids are typically vigorous and fertile in the greenhouse (Hegarty et al. 2009; Brennan et al. 2013).

Comparisons of molecular and phenotypic clines in the hybrid zone suggest that the environment is playing a role in determining trait differentiation across the hybrid zone, with selection against hybrid genotypes playing a role at more local spatial scales (Brennan et al. 2009). A nuclear gene phylogeny suggests that these are sister species within a group of closely related Mediterranean *Senecios* (Chapman MA and Filatov DA, in preparation), and they diverged very recently, probably within the last 150,000 years (11–72 thousand years ago [Ka] [Muir et al. 2013] or 130–176 Ka [Osborne et al. 2013]). Divergence with gene flow was supported (over a model without gene flow) by analysis of two transcriptome sequences (Osborne et al. 2013). Migration rates between the species are also high (Muir et al. 2013), indicating that drift alone is not strong enough to create this pattern of interspecific divergence. Coupled with the extreme environmental differences between the high- and low-altitude sites, these species are likely candidates for having arisen via ecological

speciation, that is, ecologically based divergent selection was a driving force in the speciation process (Nosil 2012).

Given that differential expression can evolve rapidly between closely related species (Whitehead and Crawford 2006; Pavey et al. 2010; Wolf et al. 2010) and can underlie species differences (Gompel et al. 2005; Abzhanov et al. 2006; Wray 2007), expression divergence could play a central role in determining phenotypic differences observed between recently derived species, even when coding sequence divergence is not apparent. In closely related species that are still exchanging genes, such as here, fixed coding differences are likely to be rare, and instead, as a complement to the expression analysis, population genetic tests can be applied to sequence data to identify loci that show signatures of divergent selection. To study the genomic basis of adaptation during ecological speciation in *Senecio*, we therefore used a next-generation sequencing (NGS) approach to 1) quantify gene expression of multiple individuals of each of the species and 2) analyze DNA polymorphism throughout the transcriptomes and compare differentially and nondifferentially expressed genes. We demonstrate that a large proportion of the *Senecio* genome is apparently able to introgress, resulting in low average interspecific differentiation, and we report evidence for differential expression and elevated differentiation in a subset of loci that are likely to be involved in adaptation to opposing selective pressures at high and low altitudes.

Results

Patterns of DNA Polymorphism

Assembly of the *Senecio* transcriptome resulted in 18,797 contigs of ≥ 500 bp. Based on BlastX comparison to the *Arabidopsis* proteome, more than half of these contigs were $>50\%$ of the full length, and almost 30% were predicted as being approximately ($>90\%$) full length (supplementary fig. S1, Supplementary Material online). After mapping reads from the 20 individuals (table 1) to the reference transcriptome and excluding loci for which more than 6 of the 20 individuals were absent, we were left with 17,394 loci, totaling 22.63 MB. The alignments for the majority of these loci (16,052; 92.3%) contained portions of all 40 alleles sequenced from all 20 individuals.

Before excluding sites with missing data, 17,250 loci were variable in the total (*S. aethnensis* plus *S. chrysanthemifolius*) data set, while 16,920 and 16,608 loci were polymorphic within *S. aethnensis* and *S. chrysanthemifolius*, respectively. In total, we identified 354,815 single nucleotide polymorphisms (SNPs), while species-specific data sets included 274,002 and 216,999 SNPs within *S. aethnensis* and *S. chrysanthemifolius*, respectively (one SNP on average every 83 and 104 bp; table 2). The overall level of sequence diversity across both species pooled (π_{both}) was 0.00361 ± 0.0038 (standard deviation [SD]), but differed significantly between the two species ($\pi_{\text{aet}} = 0.00365 \pm 0.0036$ [SD], $\pi_{\text{chr}} = 0.00290 \pm 0.0035$, paired *t*-test; $t = 23.52$; $P = 1.8 \times 10^{-119}$; table 2 and fig. 1).

Excluding sites with missing data reduced the total number of sites in the entire aligned transcriptome to 6.13

Table 1. Locations of *Senecio* Populations from Which Achenes Were Collected for This Investigation.

Code	Species	Collection Site	Latitude	Longitude	Altitude (m)	n
BRO	<i>Senecio chrysanthemifolius</i>	Bronte, Sicily	37.47°N	14.50°E	870	2
LIN	<i>S. chrysanthemifolius</i>	Linguaglossa, Sicily	37.50°N	15.07°E	750	2
NIC	<i>S. chrysanthemifolius</i>	Nicolosi, Sicily	37.37°N	15.1°E	850	4
RAN	<i>S. chrysanthemifolius</i>	Randazzo, Sicily	37.57°N	14.57°E	763	2
PRO1	<i>S. aethnensis</i>	Piano Provenzana, Sicily	37.47°N	15.1°E	2036	2
PRO2	<i>S. aethnensis</i>	Piano Provenzana, Sicily	37.47°N	15.1°E	2036	3
ET1	<i>S. aethnensis</i>	Rifugio Piccolo, Sicily	37.42°N	14.59°E	2097	4
ET3	<i>S. aethnensis</i>	Rifugio Piccolo, Sicily	37.43°N	14.59°E	2471	1
GAL	<i>S. gallicus</i>	Amoreira, Algarve, Portugal	37.21°N	8.50°W	<50	1
VER	<i>S. vernalis</i>	Cyprus	Unknown	Unknown	Unknown	1

NOTE.—Seed were collected by Adrian Brennan, with the exception of *S. gallicus*, collected by SJH, and *S. vernalis* which was donated by the Millennium Seed Bank Partnership. “n,” number of individuals sequenced.

Table 2. Summary Statistics for the Assembled Transcriptomes.

	Combined	<i>S. aethnensis</i>	<i>S. chrysanthemifolius</i>
Full data set^a			
Number of loci	17,394	17,325	17,331
Number of monomorphic loci	144	474	786
Number of alignment positions	22,595,568	22,601,091	22,604,411
Number of SNPs	354,815	274,002	216,999
$\pi \pm SD$	0.00361 ± 0.0038	0.00365 ± 0.0036	0.00290 ± 0.0035
Tajima’s $D \pm SD$	0.039 ± 0.98	-0.040 ± 0.92	-0.046 ± 1.05
Reduced data set^b			
Number of loci	8,854	10,839	9,821
Number of monomorphic loci	314	456	854
Number of alignment positions	6,131,166	8,773,007	7,054,456
Number of SNPs	87,766	107,020	67,541
$\pi \pm SD$	0.00361 ± 0.0026	0.00367 ± 0.0026	0.00286 ± 0.0025
Tajima’s $D \pm SD$	0.052 ± 0.98	-0.036 ± 0.92	-0.044 ± 1.06

^aAll loci ≥ 500 bp in length and with sequences from $\geq 70\%$ of individuals.

^bIncludes only loci with ≥ 100 gap-free alignment positions after removing sites with missing data.

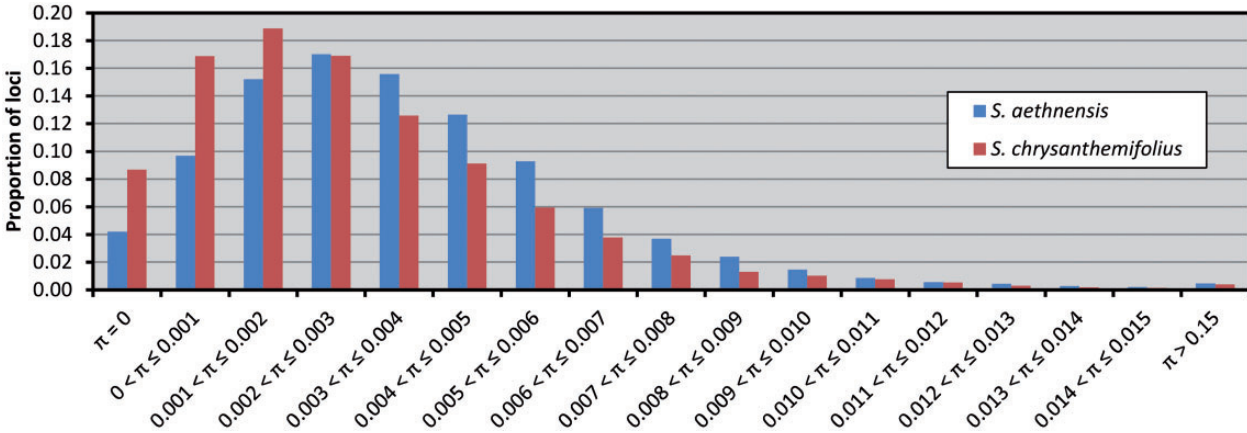


Fig. 1. Sequence diversity of the *S. aethnensis* and *S. chrysanthemifolius* transcriptomes. The distribution of sequence diversity (π , all sites) across the transcriptomes of *Senecio aethnensis* (blue) and *S. chrysanthemifolius* (red).

MB and the total number of SNPs to 87,766 (table 2). Taking each species separately, this reduced the number of SNPs from 274,002 to 107,020 within *S. aethnensis* and from 216,999 to 67,541 within *S. chrysanthemifolius* (table 2). Subsequent exclusion of loci with fewer than 100 gap-free alignment positions reduced the total data set to 8,854

alignments for the combined data set, and sequence diversity was again significantly different between the species ($\pi_{aet} = 0.00367 \pm 0.0026$ [SD], $\pi_{chr} = 0.00286 \pm 0.0025$, paired t -test; $t = 34.19$; $P = 2.4 \times 10^{-236}$; table 2). The same pattern was also found for different types of sites (noncoding, synonymous etc.; paired t -test; all $P < 10^{-5}$; fig. 2). Within species,

the same cutoff reduced the numbers of data sets to 10,839 for *S. aethnensis* and 9,821 for *S. chrysanthemifolius*. In line with the above observations of reduced polymorphism in *S. chrysanthemifolius*, the proportion of monomorphic loci was twice as high in this species (854; 8.70% of 9,821 loci) compared with *S. aethnensis* (456; 4.20% of 10,839 loci).

Across both species combined as well as within each species, Tajima's D was close to zero ($D_{\text{both}} = 0.052 \pm 0.98$; $D_{\text{aet}} = -0.036 \pm 0.92$, $D_{\text{chr}} = -0.044 \pm 1.06$). Although the distribution was not significantly different between *S. aethnensis* and *S. chrysanthemifolius* (paired t -test; $t = 1.575$; $P > 0.1$), *S. chrysanthemifolius* showed a broader spread of values (table 2 and fig. 3).

Linkage Disequilibrium

One caveat of carrying out simultaneous analysis of an entire transcriptome is that loci residing in the same region of the genome may be nonindependent in an evolutionary sense due to linkage. Positive selection at one site, for example, can affect frequencies of alleles at any sites that are in linkage disequilibrium (LD) (e.g., Maynard-Smith and Haigh 1974). To determine whether nearby SNPs were likely to be in LD with each other, we inferred the distance at which LD breaks down within the loci. Despite this analysis being based on intronless sequences and therefore the distances we used are likely to be underestimated in many cases, this is unlikely to be a very large difference as total intron length per gene in plants is on average <2 kb (Carels and Bernardi 2000; Hong et al. 2006; Jaillon et al. 2007; You et al. 2009). Above 7.5 kb, there was no evidence for pairs of SNPs to demonstrate LD; therefore, LD is only likely to affect SNPs within loci (supplementary fig. S2, Supplementary Material online), and thus the vast majority of loci are likely to be independent in evolutionary sense.

Genetic Differentiation and Demographic Inference

Almost half of the SNPs (40,094; 45.7%) were shared between the two species, and only 98 fixed differences, in only 57 loci,

were observed. Overall, genetic differentiation of *S. aethnensis* and *S. chrysanthemifolius* is quite low (for a between-species comparison), with mean and median F_{ST} values of 0.196 (± 0.191 [SD]) and 0.146, respectively (fig. 4). Nevertheless, the STRUCTURE analysis revealed $K = 2$ as the most likely number of clusters, clearly subdividing the 20 individuals into their specific groups, with only two individuals showing any evidence of admixture (fig. 5).

To infer past demography of *S. aethnensis* and *S. chrysanthemifolius*, we fit population split with migration models to the data using *dadi* (Gutenkunst et al. 2009). This approach is based on comparison of the observed and modeled site frequency spectra (SFS) (fig. 6). To distinguish the ancestral and derived alleles for each SNP, we used transcriptome sequences from two outgroup species (see Materials and Methods). To minimize the effects of selection on our demographic inference, only SNPs at 4-fold degenerate sites were used (39,635 SNPs from 4,610 loci that had at least 500 alignment positions with no missing data). We tried several models implemented in *dadi* and chose the model with the least number of parameters that was appropriate for our data; the isolation-with-migration model with population size change (*IMpre*). In this model, an ancestral population of size N_a undergoes an instantaneous size change to size N_b at time T_b . Then at time T_s it splits in two with size $s \times N_b$ and $(1 - s) \times N_b$ (where s is the fraction of N_b that forms population 1). Each is then allowed to exponentially change in size to the current effective population sizes (N_1 and N_2) and exchange migrants at rates $M_{1 \leftarrow 2}$ and $M_{2 \leftarrow 1}$ (fig. 7).

The above parameters in the model (except s) are scaled in units of $2N_a$, and thus one has to assume some value for the ancestral population size to convert into more conventional units. Using different approaches, two previous studies estimated N_a to be 227,837 (range $\sim 129,000$ – $475,000$; Muir et al. 2013) and 312,023 ($\pm 9,833$ [standard error]; Osborne et al. 2013). Assuming $N_a = 300,000$ (see Discussion for validation) and a generation time of 2 years, the time of species split is estimated to have occurred 107,571 years ago (bootstrap range [BR]: 52,853–186,785) and the contemporary

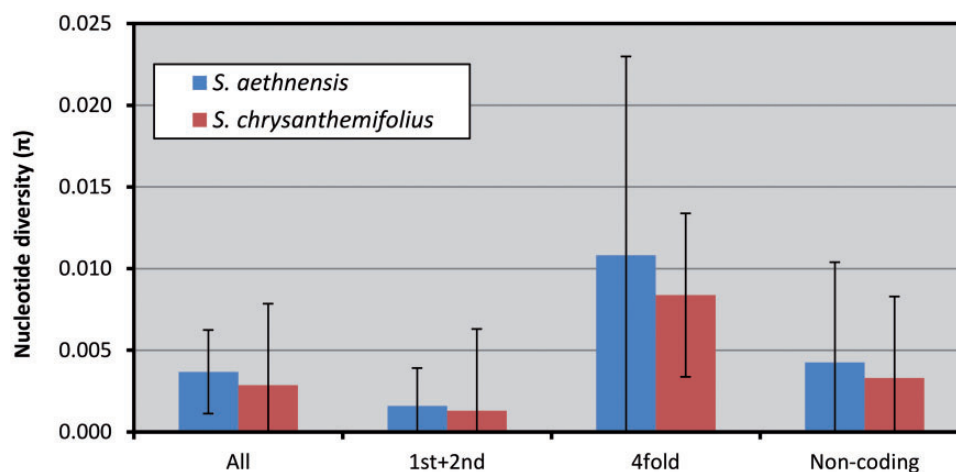


FIG. 2. Sequence diversity is significantly lower in *S. chrysanthemifolius* relative to *S. aethnensis*. Mean (\pm SD) sequence diversity per locus across all sites (All) and across groups of sites (1st and 2nd coding positions, 4-fold degenerate coding positions and noncoding positions) is given. For each test *S. chrysanthemifolius* was lower than *S. aethnensis* (paired t -test; all $P < 0.0001$).

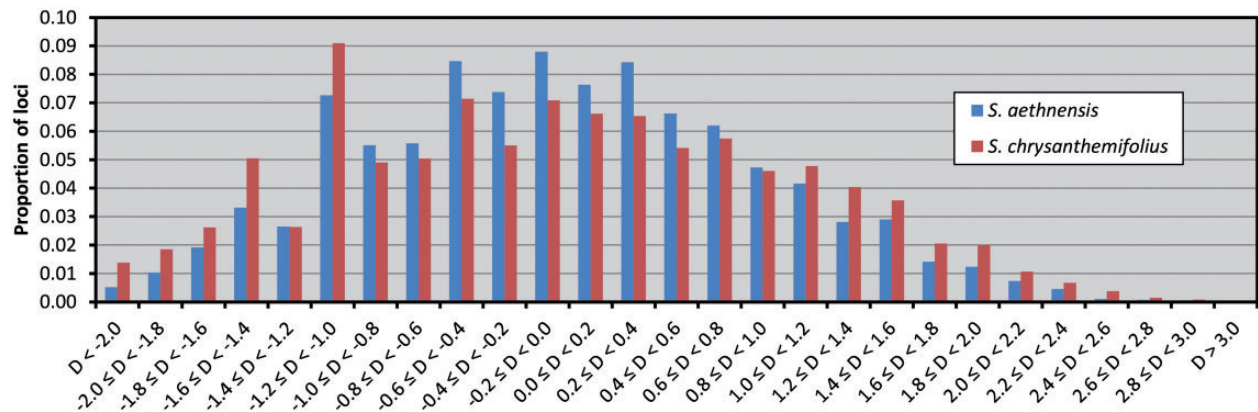


Fig. 3. Tajima's D across the transcriptomes of *S. chrysanthemifolius* shows a broader spread of values than in *S. aethnensis*. The proportion of loci that fall within each bin is given per species. Mean Tajima's D was not significantly different between the species (paired *t*-test; *P* > 0.1).

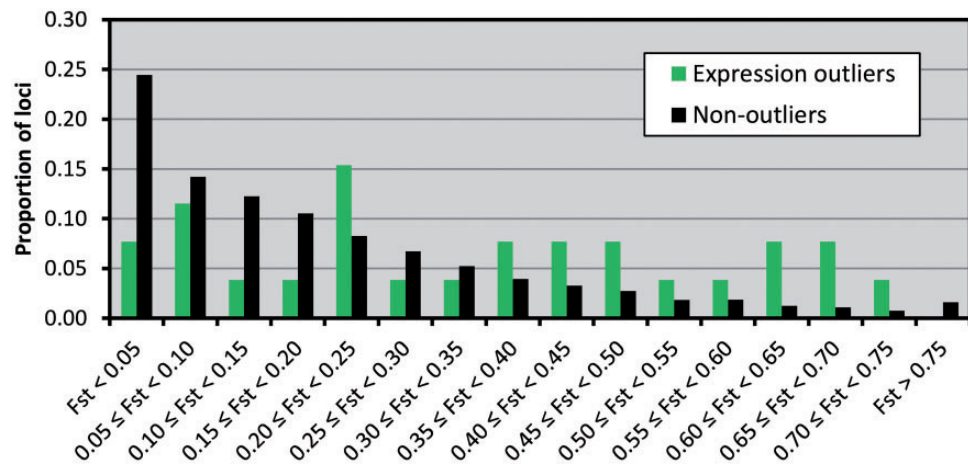


Fig. 4. F_{ST} is strongly skewed toward low values across all loci; however expression outliers showed a greater proportion of loci with high F_{ST} . The expression outliers with sufficient sequence information ($n = 26$ loci; see text) are plotted in green and nonoutliers in black.

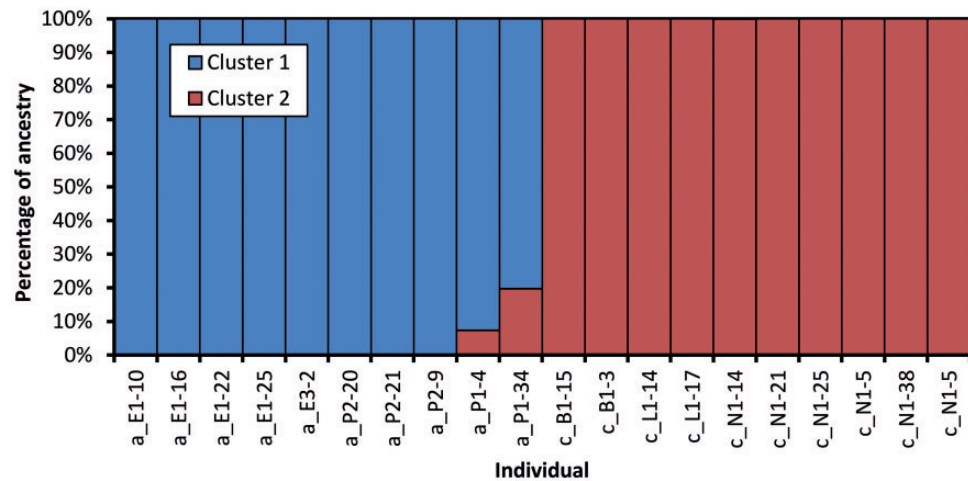


Fig. 5. *Senecio aethnensis* and *S. chrysanthemifolius* are genetically differentiated. STRUCTURE was used to determine that the most likely number of genetic clusters was 2 (blue and red). All *S. chrysanthemifolius* individuals ("c_xxx") fall into cluster 2 with >99.9% ancestry; however, for two *S. aethnensis* individuals ("a_xxx") slight (<20%) mixed ancestry was uncovered. Individuals are named according to Table 1.

population sizes of *S. aethnensis* and *S. chrysanthemifolius* are 54,027 (BR: 26,448–89,856) and 39,846 (BR: 19,987–63,642), respectively (fig. 7). There appears to be a slight overall reduction of population size compared with the ancestral species,

which was also reported by Muir et al (2013). This is due to the reduction in population size of *S. aethnensis* because *S. chrysanthemifolius* has apparently shown an increase in population size recently (fig. 7). Interestingly, the *IMpre* model

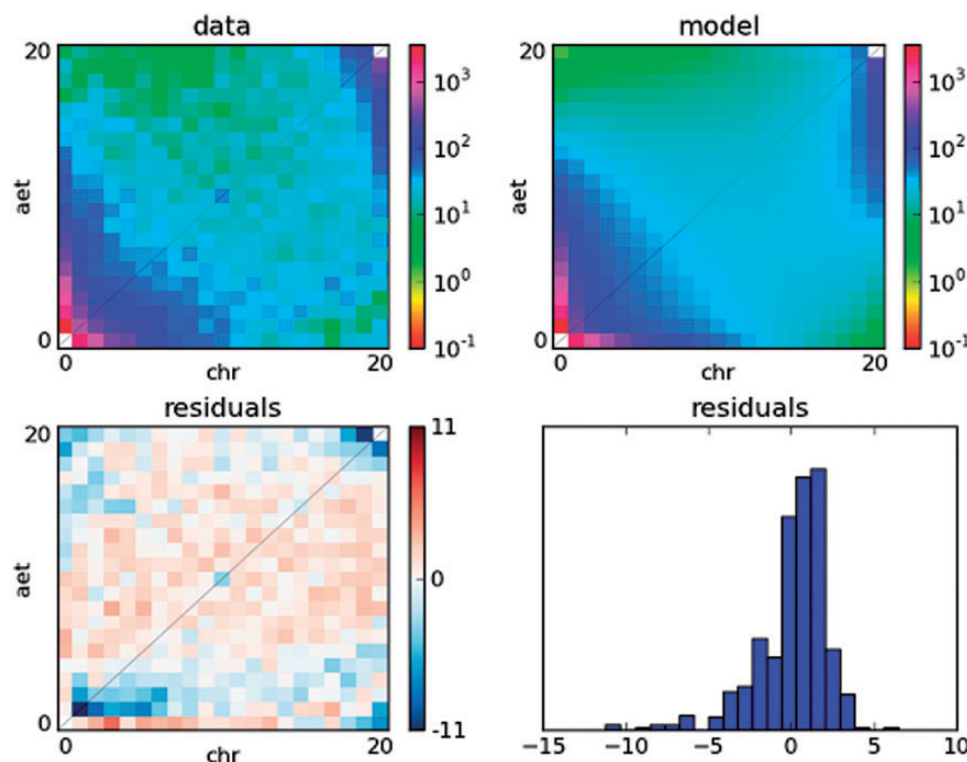


FIG. 6. The fit of the *IMpre* demographic model to the data summarized by the observed and modeled 2D-SFS. (A and B) Observed (A) and modeled (B) 2-D SFS for *S. aethnensis* (Y axis) and *S. chrysanthemifolius* (X axis). The scale on the right indicates the number of polymorphisms in each frequency class (squares of the grid). (C) Anscombe residuals between the model and the data where blue and red (negative and positive values) indicate the model predicts too few and too many polymorphisms in a given grid square, respectively. (D) The distribution of the residuals.

indicates a very uneven (94% to 6%) split of the ancestral population at time T_s into *S. aethnensis* and *S. chrysanthemifolius* (fig. 7), which suggests that the latter originally was founded as a small population at the margin of the species range.

Gene flow from *S. aethnensis* to *S. chrysanthemifolius* ($M_{2 \leftarrow 1} = 9.64$ [BR: 5.35–16.81]) appears to be higher than in the opposite direction ($M_{1 \leftarrow 2} = 4.94$ [BR: 2.70–9.09]). Overall, migration rate in both directions is quite high and is significantly different from zero; forcing the migration parameters to equal zero dramatically reduces the likelihood of the model (model with bidirectional migration log likelihood $\ln L = -2,157.64$; models with migration in only one direction $\ln L = -13,756.64$ and $-3,533.61$; model with no migration $\ln L = -21,738.14$; likelihood ratio test for all comparisons $P < 0.001$). This high migration rate is not unexpected, given the field observations of extensive hybrid swarms at intermediate altitudes (e.g., Ross 2010 and personal observations).

Global Selective Sweeps

A mutation that is adaptive across the entire range of populations interconnected by gene flow is expected to spread throughout the species range, even if gene flow between populations (or species/subspecies/races) is very low (Slatkin and Wiehe 1998). Unlike diversifying selection and local adaptation, “global sweeps” are expected to eliminate genetic differentiation between the populations or species. Such sweeps are expected to leave distinctive footprints in

DNA polymorphism, such as reduced variation, biased frequency spectra, and elevated LD across the entire range of hybridizing populations (Oleksyk et al. 2010).

Comparative analyses of the patterns of polymorphism and deviations from neutral allele frequency spectra, as well as explicit tests for nonneutrality, are therefore instructive in this regard. First, we identified 314 loci (3.5% of 8,854) that contained no polymorphism in the sequences from both species. To test whether low diversity in these loci is incompatible with neutrality, we used a maximum likelihood version of the Hudson-Kreitman-Aguade (HKA; Hudson et al. 1987) test (Wright and Charlesworth 2004), integrated in Proseq (Filatov 2009). Maximum likelihood (ML)-HKA tests rejected neutrality ($P < 0.05$) for 22 of the loci (7.0%). Second, we investigated the patterns of polymorphism in the (nonmonomorphic) loci in the lowest 5% tail of the distribution of average heterozygosity (π). Average π in this subset of loci was only 0.00034 ± 0.00015 (SD). Eighty eight (>20%) of these had among the lowest (most negative 5%) Tajima's D , suggestive of recent selection, and eight were in the lowest 5% of Fay and Wu's H distribution, specifically indicative of recent (i.e., <0.1 Ne; Przeworski 2002) positive selection at these loci. Taken together, these analyses suggest that very low or zero polymorphism in a subset of loci is due to recent selection.

Diversifying Selection

Diversifying selection at a locus is expected to increase differentiation and inflate F_{ST} . This signal is used by the BayeScan

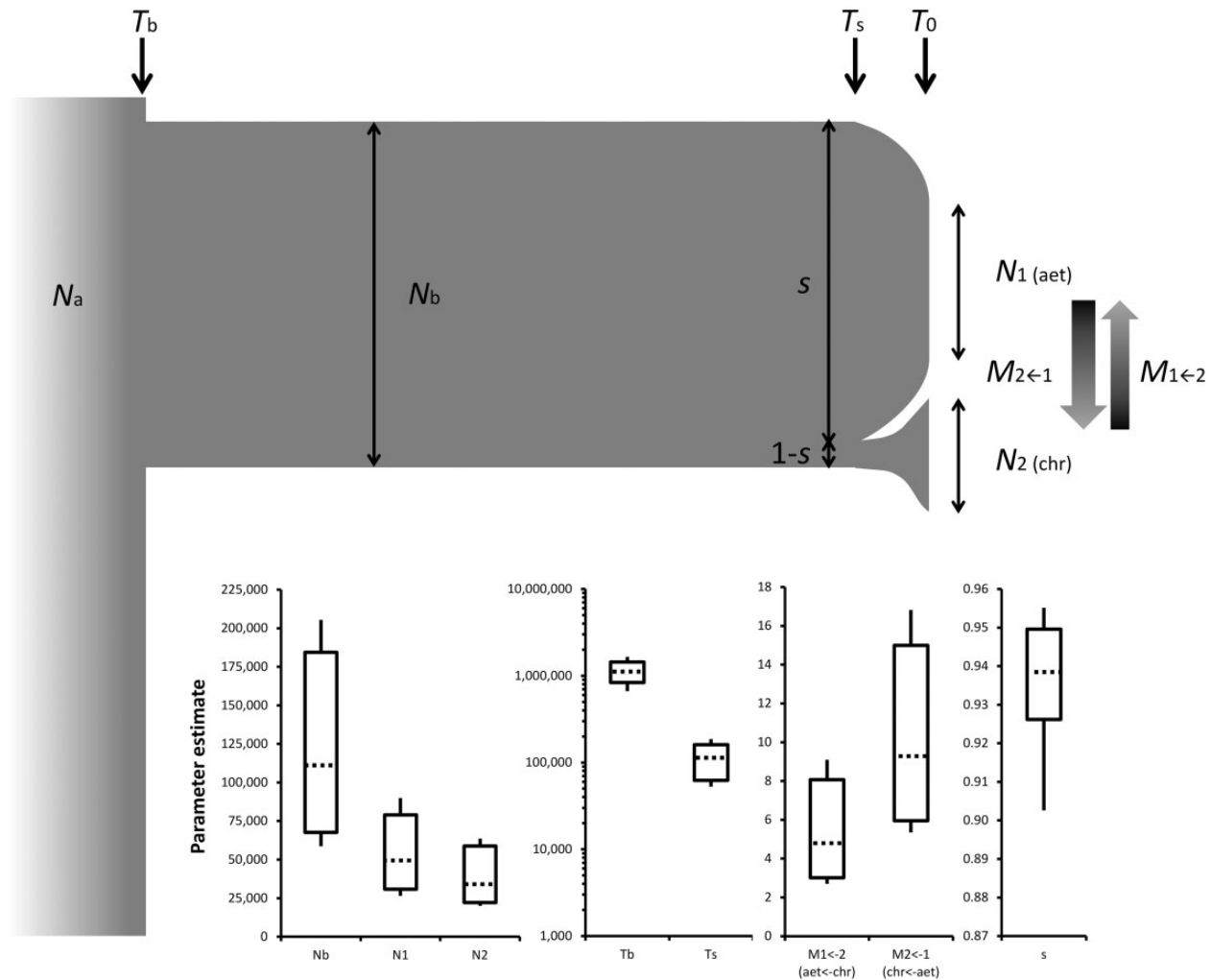


FIG. 7. Graphical summary of the most likely inferred demographic scenario of speciation. This scenario was inferred with the *IMpre* model implemented in *dadi*. The parameters are described in the text. Time (X axis) and each population size (Y axis) are drawn to scale based on the mean parameter estimates. (Inset) Box and whisker plots of the parameter estimates from 300 bootstrap replicates. Boxes indicate the 95% confidence intervals, whiskers indicate the range, and the dashed line indicates the mean estimate. Note the log₁₀ scale for the parameters T_b and T_s .

software (Foll and Gaggiotti 2008; Fischer et al. 2011), which identified 199 positive outlier loci (i.e., greater than expected differentiation) with 1% false discovery rate (FDR). Although BayeScan does not simply identify the loci with the highest F_{ST} , average F_{ST} for these positive outliers was 0.657 ± 0.193 (SD), significantly greater than the remainder of the loci (Mann–Whitney $U = 8.15 \times 10^4$, $z = -21.77$, $P = 4.02 \times 10^{-105}$). Average divergence, D_{xy} , was also significantly higher in the BayeScan outliers (0.0044 ± 0.0026 [SD] vs. 0.0034 ± 0.0022 ; Mann–Whitney $U = 5.876 \times 10^5$, $z = -6.278$, $P = 3.43 \times 10^{-10}$) suggestive of selection driving divergence in this subset of loci. Out of 98 SNPs fixed between the two species, 49 were found in the 199 BayeScan outlier loci.

An additional 369 loci were deemed negative outliers, which may contain a subset of loci with the signature of balancing selection. These loci had a significantly lower F_{ST} than the remainder of the loci (0.099 ± 0.092 [SD] vs. 0.200 ± 0.193 ; Mann–Whitney $U = 1.059 \times 10^6$, $z = -9.678$, $P = 3.74 \times 10^{-22}$).

Within the positive outliers are a number of loci with specific signatures of diversifying selection or species-specific adaptation. Average heterozygosity (π) for the BayeScan outliers was significantly lower than the nonoutlier loci in *S. chrysanthemifolius* (0.00060 ± 0.0009 [SD] vs. 0.00290 ± 0.0025 ; Mann–Whitney $U = 2.572 \times 10^5$, $z = -16.94$, $P = 2.10 \times 10^{-64}$) and marginally so in *S. aethnensis* ($\pi = 0.00318 \pm 0.0022$ [SD] vs. 0.00368 ± 0.0026 ; Mann–Whitney $U = 7.669 \times 10^5$, $z = -2.643$, $P = 0.0082$). This reduction in the level of polymorphism (when each species was compared with that expected from divergence from out-group *S. vernalis*) was significant in 35 loci (ML-HKA test, $P < 0.05$): 21 in *S. chrysanthemifolius*, 11 in *S. aethnensis*, and 3 loci in both species. Moreover, a complete absence of polymorphism was found for 68 of the 199 loci (34.2%) in *S. chrysanthemifolius*, significantly more than expected from the genomewide pattern (Fisher's exact test $P = 8.690 \times 10^{-11}$), while only eight (4.0%) were monomorphic in *S. aethnensis* (Fisher's exact test $P = 0.325$). Three of the 199 outlier loci were monomorphic in both species (with one to three

SNPs differentiating the alleles). The loci showing elevated differentiation coupled with reduced polymorphism in one or both species represent some of the strongest candidates for diversifying selection or parallel adaptation.

Species Differentiation in Gene Expression

Expression (reads per kilobase of exon model per million mapped reads [RPKM]) was calculated for each of 18,797 contigs in 20 separate individuals, and *t*-tests were used to identify significantly different expression between species (see Materials and Methods; [supplementary table S1, Supplementary Material](#) online). With an FDR value of 5%, significantly different expression was detected for 64 contigs, with 29 and 35 being expressed at a significantly higher level in *S. aethnensis* and *S. chrysanthemifolius*, respectively. The NOISeq test (Tarazona et al. 2011) for differential expression revealed a smaller number of gene candidates (32), with 9 expressed at a higher level in *S. aethnensis* and 23 higher in *S. chrysanthemifolius*. Six loci were detected as showing significant expression differences in both the *t*-test and using NOISeq, giving a total of 90 expression candidates. The fold difference of the expression candidates between *S. aethnensis* and *S. chrysanthemifolius* was sometimes extreme, $>100\times$ in 11 cases (see [supplementary table S1, Supplementary Material](#) online).

There was little overlap for the differentially expressed loci identified by this study and the previous microarray investigation (Hegarty et al. 2009); of the 100 differentially expressed transcripts identified by Hegarty et al. that were identified in our transcriptome, only two (see [supplementary table S1, Supplementary Material](#) online) were differentially expressed in our RNA-Seq investigation. The Hegarty et al. (2009) investigation was based on a different tissue mix (two different bud stages only) and pooled samples from the different taxa, and hence represents more individuals per species but in the absence of biological replicates (see Discussion). Thus, the two analyses are not strictly comparable.

The genes differentially expressed in *S. aethnensis* and *S. chrysanthemifolius* showed evidence for higher genetic differentiation at the sequence level relative to the remainder of the loci. Initially, divergence estimates (F_{ST} and D_{xy}) were available for only 17 of the expression candidates because these differentially expressed genes often have very low coverage (and hence a lot of missing data) in one of the species. We therefore polymerase chain reaction (PCR) amplified and Sanger-sequenced a further nine expression outlier loci from genomic DNA from the same individuals. F_{ST} between *S. aethnensis* and *S. chrysanthemifolius* was significantly greater for the 26 expression outliers than the nonoutliers (0.348 ± 0.226 [SD] vs. 0.195 ± 0.191 ; Mann–Whitney $U = 6.594 \times 10^4$, $z = -3.57$, $P = 0.0004$; [fig. 4](#)). This was also the case when comparing just the initial 17 loci to the nonoutliers ($F_{ST} = 0.344 \pm 0.242$; $P = 0.0081$) as well as the nine loci amplified by PCR ($F_{ST} = 0.355 \pm 0.205$; $P = 0.0160$). Similarly, D_{xy} , the average number of pairwise differences between the alleles from the two species, was significantly higher for the differentially expressed loci (0.0070 ± 0.0051 [SD] vs.

0.0035 ± 0.0022 ; Mann–Whitney $U = 6.015 \times 10^4$, $z = -4.031$, $P = 5.547 \times 10^{-5}$).

Despite this trend, only one of the expression candidates (a DNAJ heat-shock protein homologue) was also a positive BayeScan-based candidate (i.e., outlier with increased inter-specific divergence); however, only 17 of the expression outliers had sufficient sequence information to be included in the BayeScan analysis. This and a further three of the 17 expression candidates exhibited an F_{ST} in the top 5% of the transcriptomic distribution, comprising a serine/threonine kinase, a methyltransferase, and a pre-mRNA-splicing factor.

Across both species combined, sequence diversity (π) for the expression outliers was higher than for the nonoutliers (Mann–Whitney $U = 5.476 \times 10^{-4}$, $z = -2.29$, $P = 0.02223$), but not so for the within species diversity (all $P > 0.1$). Tajima's *D* was also not significantly different for the expression outliers (all $P > 0.1$); however, it should be noted that this is based on a small number of outliers.

In addition, ML-HKA tests were conducted for each species separately for the expression candidates that also had a *S. vernalis* sequence. For the species-specific samples, the number of differentially expressed loci with ≥ 100 gap-free sites increased (compared with the two species sample) to 19 for *S. aethnensis* and 30 for *S. chrysanthemifolius*. For eight loci, the ML-HKA was significant (three loci in *S. aethnensis*, four loci in *S. chrysanthemifolius*, and one locus in both species), suggesting that directional selection for differential expression had taken place in these species at these loci.

Types of Loci under Selection

Analysis of gene ontology (GO) terms revealed significant ($P = 0.00048$) overrepresentation of genes involved in trehalose biosynthesis among the positive BayeScan outliers, i.e., those showing the signature of adaptive differentiation. Trehalose is a disaccharide that plays a role in desiccation tolerance in bacteria and yeast (Crowe et al. 1992), and, although trehalose is not thought to accumulate in the majority of plants (the exception being “resurrection plants” (e.g., [Vicre et al. 2004](#)), it may be that genes involved in trehalose biosynthesis are under selection at the sequence level due to differing drought conditions of the two habitats. Using an FDR of 5%, this GO term was, however, not significant.

GO annotation of the expression candidates revealed an overrepresentation of genes involved in primary amine oxidase activity ($P = 0.00057$) and magnesium ion binding ($P = 0.00680$), but these again were not significant when a 5% FDR was used. The only differentially expressed gene that was also among the positive Bayescan outliers (contig 14541) encoded a DNAJ heat shock N-terminal domain-containing protein, thought to be involved in response to heat stress, which may reflect adaptation of low-altitude *S. chrysanthemifolius* to hotter, more arid conditions at the bottom of Mt. Etna. Functional annotation of further differentially expressed genes with high F_{ST} (in top 5%) or reduced DNA polymorphism (significant ML-HKA) provided no obvious connections to adaptations to altitude.

Discussion

During speciation, a heterogeneous pattern of genomic differentiation can build up in response to divergent selection of adaptive alleles at a handful of loci, while the rest of the genome is largely free to exchange alleles (Wu 2001). This heterogeneous pattern may become more prominent if the recently diverged species come back into contact and are free to exchange alleles throughout the genome, while maintaining their interspecific differences due to the reduced or prevented introgression of locally adaptive alleles (Payseur 2010). There are only a few examples of analyses of clear-cut ecological speciation with ongoing gene flow (reviewed in Funk 2009), and the majority of research has been carried out on animal systems (Nosil 2007; Schwarz et al. 2009; Renaut et al. 2010) and hardly on plants (however, see e.g., Schemske and Bradshaw 1999; Savolainen et al. 2006; Papadopoulos et al. 2011 for plant systems). Such focus on animals in this field is somewhat surprising, as the pervasiveness and importance of gene flow between closely related species has long been recognized in the plant literature (e.g., Anderson 1949; Anderson and Stebbins 1954; Stebbins 1959; reviewed in Rieseberg 1995; Arnold 1997).

Ecological Speciation in *Senecio*

We undertook a transcriptomewide analysis of polymorphism and divergence in a recently derived pair of plant species, which are thought to have diverged through ecological speciation. By comparing genomewide and locus-specific patterns of sequence variation and gene expression, we can begin to understand how these species formed and/or how they are being maintained in the face of ongoing interspecific gene flow. The study system herein concerns the sister species *Senecio aethnensis* and *S. chrysanthemifolius* (Asteraceae) that form a hybrid zone on Mt. Etna, Sicily. Contrasting environments of high and low altitudes create strong divergent selective pressures in geographically proximal (within a few kilometer) populations of these species.

Previous demographic inferences indicate a relatively recent (~35–150 Ka) split of *S. aethnensis* and *S. chrysanthemifolius*, a smaller population size in the latter species, and extensive interspecific gene flow (Muir et al. 2013; Osborne et al. 2013). The genomewide data reported in the current study is consistent with this scenario, with our estimate of the time of species split (~108 Ka) falling in between the two previous estimates. Our relatively high estimate of interspecific gene flow demonstrates that the gene pools of the two species are actively mixing and that without diversifying selection at genes involved in adaptation to contrasting environments, the distinctiveness of these species would be wiped out by ongoing gene flow (Wright 1931; Slatkin 1985).

Our estimates of the split time and population sizes are based on the assumption of an ancestral population size, N_a , and assuming smaller or larger values would proportionately reduce or increase the estimates. $N_a = 300,000$ was based on the previous literature (Muir et al. 2013; Osborne et al. 2013), and furthermore, this estimate gives reasonable values for the mutation rate, μ , calculated from the scaled mutation rate, θ

($=4N_a\mu$), estimated in this model. Assuming $N_a = 300,000$ results in an estimate of μ (4.27×10^{-9} ; bootstrap range, BR: 2.42×10^{-9} to 7.73×10^{-9}), which is roughly compatible with an Asteraceae-specific mutation rate ($\mu \sim 1 \times 10^{-8}$; Strasburg and Rieseberg 2008) and a generic plant mutation rate ($\mu \sim 5 \times 10^{-9}$ Wolfe et al. 1987).

Our demographic inference using *dadi* confirmed the previous conclusion that *S. aethnensis* has a somewhat larger effective population size compared with *S. chrysanthemifolius* (Muir et al. 2013), which is also in line with various summary statistics reported here (table 2). In particular, DNA polymorphism in *S. aethnensis* was significantly higher than in *S. chrysanthemifolius* (figs. 1 and 2), and the former species contained less monomorphic loci (4.2%) than the latter (8.7%). This may seem somewhat surprising given the wider extant distribution of *S. chrysanthemifolius* compared with *S. aethnensis*, which is restricted to the upper part of Mt. Etna; however, the past distributions and population sizes of the two species may have been very different. First, the split of the ancestral population at time T_s gave rise to an ancestral *S. aethnensis* population many times larger than that of *S. chrysanthemifolius* (fig. 7); therefore, it seems likely that the now relatively widespread *S. chrysanthemifolius* originated as a small marginal population. Second, because *S. aethnensis* is better adapted to colder conditions, it is likely to have been much more widespread during the glacial cycles of the Pleistocene, whereas *S. chrysanthemifolius* may have been more restricted in its distribution compared to now, as seen in other Sicilian plant taxa (Sadori et al. 2008). Since the beginning of the Holocene ~12,000 years ago, *S. aethnensis* is likely to have become restricted to its current high altitude refuge on Mt. Etna, and *S. chrysanthemifolius* may have been able to colonize a larger area of Sicily. This scenario is in line with the overall reduction in population size of *S. aethnensis* and the increase in population size of *S. chrysanthemifolius* since their split around 108,000 years ago (fig. 7).

A Heterogeneous Pattern of Genomic Differentiation

In addition to the proximity of two species, factors such as seed dispersal by wind, the constant availability of empty niches due to volcanic activity, and an apparent lack of any reproductive barriers (Chapman et al. 2005) should provide ample opportunities for interspecific hybridization and homogenization of the genomes of *S. aethnensis* and *S. chrysanthemifolius*. Hybrid populations flourishing at intermediate altitudes may also act as a bridge (or rather stepping stones) for interspecific gene flow (Brennan et al. 2009). In this respect, our findings of relatively low genomewide genetic differentiation (only 98 fixed SNPs in ~10,000 genes analyzed), extensive sharing of SNPs (45% of all SNPs were shared across both species) and a high migration parameter ($2N_a m = 4.94$ – 9.64) are not entirely unexpected. In addition, a small number of loci potentially show the signature of a global selective sweep in which an allele has risen to fixation (or almost so) across both species. Morjan and Rieseberg (2004) suggested that the collective evolution of species exchanging adaptive

alleles may be fairly widespread, but, at the time, molecular genetic data from nuclear protein coding genes supporting this was lacking. Our data, with caveats (see later), suggest the presence of global (transspecific) sweeps in these two species.

Despite extensive gene flow, the species, however, maintain their phenotypic (and genotypic; [fig. 5](#)) integrity at the extremes of the transect, and therefore a portion of the genomes is presumably being prevented from introgressing (in contrast to the remainder of the genome), due to strong, most likely ecological-based, selection. Indeed, this is supported by our data, with a subset of loci showing significantly elevated differentiation. The pattern of polymorphism across the transcriptome sheds light on the process of ecological speciation.

It appears that only a small subset of the genome harbors evidence for playing a role in the specific differences between these species, with only 90 loci (of ~18,800) showing gene expression divergence and 199 loci showing significantly elevated differentiation (at $FDR < 0.01$). This fraction of sequence-based outliers (199/8,540; 2.33%) is low compared with other studies (reviewed in [Strasburg et al. 2012](#)); however, comparisons between studies are difficult to make because of differences in methodologies and significant cutoffs ([Nosil et al. 2009](#)). For example, our study utilized coding region polymorphisms, whereas most related studies use anonymous markers ([Strasburg et al. 2012](#)). Similarly, our study used a very large number of loci at the expense of the number of individuals analyzed. It remains to be seen whether other such large-scale investigations of coding region polymorphism give similar proportions of outlier loci.

Evolution of Gene Expression

The analysis of gene expression under controlled conditions revealed a handful of loci differentially expressed between the two species, which is consistent with the view that gene expression differences can evolve quickly between species or differentially adapted populations ([Whitehead and Crawford 2006](#); [Wolf et al. 2010](#)). Evolution of such loci would require fixation of *cis*-mutations in the regulatory regions and/or evolution of a gene that acts in *trans*. Under *cis*-regulatory evolution, one would expect to detect higher differentiation as well as signatures of selective sweeps in the 5' regions of these loci. Because of the nature of RNA-seq, our data did not allow us to investigate the upstream regions of these loci, and, as mentioned previously, differentially expressed genes usually had very low expression in one of the species. Despite this, the initial 17 expression outliers with sufficient sequence data in both species did show increased interspecific differentiation (F_{ST} and D_{xy}) relative to the genome as a whole, and this was corroborated when a further nine outlier loci were PCR amplified and sequenced ([fig. 4](#)). This was also found when six differentially expressed loci (identified in [Hegarty et al. 2009](#)) were compared with 11 genes with no evidence of differential expression ([Muir et al. 2013](#)). This pattern could be due to adaptive divergence in a proximal region that is responsible for regulating gene expression (promoters or nearby enhancers that were not

sampled using RNA-seq) and a parallel increase in differentiation of the coding region due to linkage to the adaptive mutations. A subset of the expression candidates did show departure from neutrality in the coding regions in the ML-HKA test in one or other species, and this ad hoc test therefore informs us about which species is likely to have experienced the selection for altered expression.

GO terms related to primary amine oxidase activity and magnesium ion binding appear to be overrepresented among the differentially expressed genes. It is possible that differential metal ion deposition through volcanic activity (e.g., [Aiuppa et al. 2006](#)) is imposing a selection pressure, although, at this stage, it is too early to say that this is the reason for overrepresentation of loci involved in ion binding.

Only two loci were found to be differentially expressed in both our study and a previous microarray analysis of gene expression differences between these two species of *Senecio* ([Hegarty et al. 2009](#)). This is probably due to differences in tissues analyzed and methodology. [Hegarty et al. \(2009\)](#) used RNA extracted from capitulum buds and flower buds, whereas our analyses used RNA extracted from capitulum buds, leaf, and stem tissue; thus, the two analyses are not strictly comparable, so it is perhaps not surprising that more loci in common were not identified. Further, [Hegarty et al.](#) did not seek to explore intraspecific variation in gene expression, but rather created an “average” transcriptome by pooling RNA from multiple individuals to identify gross species-specific differences between the two species. Such an approach without individual-level analysis could lead to the identification of false positives, for example, if a gene was very actively expressed in only one individual out of the pool, its transcript abundance would appear very different in the two species. Our NGS RNA-seq analysis overcomes this issue because we control for variation in expression among individuals within the species. Thus, the two loci identified as being differentially expressed by both microarray and RNA-seq are particularly interesting together with the other candidate transcripts identified by RNA-seq analysis. These 90 candidate genes will be the subject of future studies of the evolution of differential gene expression in *Senecio*.

Caveats and Cautions

A study of this magnitude may be affected by various types of errors, and it is important to consider their sources and possible effects on our conclusions. Errors in both sequencing and read mapping may result in an excess of low-frequency variants in the data, which would lead to biases in demographic inferences. There is no evidence of an excess of low-frequency variants in our data, if anything there may be a slight lack of such sites (see [supplementary fig. S3, Supplementary Material online](#)). If, however, the rare variants were clustered in just a few genes, then this pattern of polymorphism may mimic recent selective sweeps, which could partly account for the signature of transspecific selective sweeps we detected in the data.

LD is heterogeneous across the genome (e.g., [Kim et al. 2007](#)), and a localized genomic region of increased LD would cause hitchhiking of many genes linked to target of selection;

however, this is not anticipated to have a major influence on the data we present. As we demonstrated earlier, the extent of LD in the genomes of these *Senecios* is low (supplementary fig. S2, Supplementary Material online), with LD likely to be limited to SNPs within loci and not expected to extend to neighboring loci.

Finally, we acknowledge that transcriptomics only captures the transcribed portions of the genome that are expressed at a sufficient level in the tissues we chose to analyze. Future analyses aimed at analyzing the signature of selection throughout the remainder of the genome will shed light on the contribution of selection on coding versus regulatory DNA during speciation and adaptation.

Conclusions

This study represents one of the first transcriptomewide analyses of adaptation to contrasting environments during ecological speciation in plants. Our results suggest that the evolution of morphological and ecological differences between the two *Senecio* species on Mt. Etna occurred over a relatively short timescale, and it was accompanied by diversifying or species-specific selection in a small proportion (<3%) of the sequenced loci. Selection at this small portion of the genome is apparently sufficient for the species to maintain their phenotypic and genotypic differences on Mt. Etna, despite a high level of ongoing gene flow that has the potential to homogenize the rest of the genome.

Materials and Methods

Study Species, RNA Extraction, and Sequencing

Achenes (single-seeded fruit) of *S. aethnensis* and *S. chrysanthemifolius* were collected on Mt. Etna from wild plants at the extremes of the altitudinal cline to ensure the best approximation of species genetic “purity” (table 1). All *S. aethnensis* achenes were collected from above 2,000 m. within ~4 km of the summit, and achenes of *S. chrysanthemifolius* were taken from a range of populations at the base of Mt. Etna (all ~700–800 m above sea level [a.s.l.]) ~12–15 km from the summit to the north, west, south, and northeast. Achenes of the outgroups (based on Comes and Abbott 2001) *S. vernalis* and *S. gallicus*, respectively, were supplied by the Millennium Seed Bank Partnership (<http://data.kew.org/seedlist/index.html>, last accessed October 9, 2013) and collected by S.J. Hiscock. Achenes were germinated on damp filter paper and seedlings transferred to a soil/vermiculite mix in identical-sized pots in a growth room set at 19–21 °C with a 16-h photoperiod. Ten plants of each focal species and a single outgroup were selected, and RNA was extracted from the plants as the first inflorescence opened. The apical tissues (comprising inflorescence, stem, and uppermost leaf) were removed from the plants and immediately frozen in liquid nitrogen prior to RNA extraction with the RNeasy kit (Qiagen, Crawley, UK). The libraries for high-throughput sequencing of the polyA fraction of the transcriptome were prepared and sequenced at the Genomics facility at the Wellcome Trust Centre for Human Genetics (WTCHG), Oxford. All sequencing was conducted on HiSeq2000

Illumina machines with multiplexing several libraries per lane. At least 12 million 50 bp end reads were generated for each of the libraries (see supplementary table S2, Supplementary Material online, for details).

Assembly and Annotation

Our data processing and assembly strategy were similar to that we used in our previous work (Chibalina and Filatov 2011). Illumina reads for each individual transcriptome were imported into CLC Genomics Workbench (CLC-GW; ver. 5; <http://www.clcbio.com>, last accessed October 9, 2013), trimmed of adapters and filtered for low-quality reads with default settings. Because we were interested in differential expression, it is possible that some of the most interesting genes would be unaccounted for in the sequence reads of a single individual (if expression was very low or zero in some individuals). We therefore assembled a “hybrid transcriptome,” comprising reads from one individual of each species. This transcriptome was only used as a reference for mapping short reads from each of the transcriptomes, and it was not used in any of the DNA polymorphism/divergence analyses described below. This reference *Senecio* transcriptome was assembled de novo in CLC-GW with high stringency to reduce the likelihood of paralogous transcripts assembling as a single transcript (mismatch cost 3, insertion cost 3, deletion cost 3, minimum identity 0.95). The assembly process included a scaffolding step based on paired end reads. Contigs less than 500 bp were not considered further.

The coding regions and associated GO terms for individual cDNA contigs were identified using Blast2GO (Conesa et al. 2005; Gotz et al. 2008). The XML output from Blast2GO was imported to ProSeq (Filatov 2009) and assigned to the reference transcriptome as well as to all the sequences in the alignments (described later). The scaffolding stage of the assembly resulted in short stretches of unknown sequence in places. The length of such regions was estimated by the scaffolding algorithm in CLC-GW based on distance between paired reads. Occasionally, these poly(N) regions caused a frameshift in the open reading frame (ORF), which was detected and resolved in ProSeq (Filatov 2009) by adding additional Ns to restore the ORF. At this stage, a small number of contigs with nonplant top BLAST hits and/or more than one ORF were removed, resulting in the final set of reference cDNA contigs. The completeness of the reference assembly was checked by BlastX comparison of each contig to the TAIR10 (The *Arabidopsis* Information Resource) *Arabidopsis* proteome using BLAST+ (ver. BLAST2.2.25+; Camacho et al. 2009).

Reads from each of the 20 individuals (plus the outgroup) were mapped onto the reference in CLC-GW, one library at a time, to produce consensus contigs. The same stringency settings were used as for the initial assembly. The files were exported from CLC-GW as BAM files and, using a custom perl script (available on request), parsed through samtools (Li et al. 2009) using the commands “mpileup,” “vcf2fq,” and “fq2fa” to align the reads to the reference, create consensus (with heterozygous bases encoded with IUPAC code) for each locus,

and to save sequences in fasta format. Minimum base quality was set to 20, minimum mapping quality to 5, and minimum read depth to 5. The fasta files from *samtools* were imported into Proseq (Filatov 2009), aligned to the reference, and coding regions copied across to all alleles. Heterozygous bases were phased by exporting each alignment as a *.inp file, running each alignment through fastPHASE (Scheet and Stephens 2006) using a custom perl script, and importing the resultant phased *.out files back into Proseq.

Population Genetic Analyses and Identification of Sequence-Based Outlier Loci

To investigate the clustering of genetic variation between individuals, we concatenated all SNPs from the unphased data that were present in all 20 individuals. The program STRUCTURE (Falush et al. 2003) was used to assess the number of genetic clusters (K) apparent in the data. Four runs for each K were carried out for K = 1–6, and the Evanno et al. (2005) method (implemented in *Structure Harvester*; Earl and von Holdt 2012) was used to determine the maximal value of delta(K) and hence the most likely value of K.

To determine whether LD between loci was likely to be affecting locus-specific patterns of polymorphism and divergence due to background selection, we investigated the relationship between inter-SNP distance and LD. In the absence of the genome sequence, we were limited to intronless data, so the distances between SNPs are underestimates in most cases, as the length of introns is not taken into account (see Results). χ^2 tests of independence were carried out for all pairs of SNPs in loci with at least 1,000 gap-free alignment sites ($n = 1,817$ loci) and binned according to distance between SNPs. Within each bin, we then counted the proportion of significant ($P < 0.001$) χ^2 tests (where a significant value indicated departure from linkage equilibrium; [supplementary fig. S2, Supplementary Material](#) online).

For the population genetic analyses, sites with missing data were excluded, as were loci with less than 28 *S. aethnensis* and *S. chrysanthemifolius* alleles (i.e., 14 individuals). Transcriptome-wide distributions of Nei and Li's π (Nei and Li 1979), a measure of nucleotide diversity, and Tajima's D (Tajima 1989), a measure of neutrality, were calculated in Proseq (Filatov 2009) and compared between species. Wright's F_{ST} (Wright 1951), a measure of genetic differentiation, and absolute and net divergence (D_{xy} and D_{ai} ; (Nei 1987) were calculated between *S. aethnensis* and *S. chrysanthemifolius* also using Proseq. Fay and Wu's H (Fay and Wu 2000) was calculated in DnaSP (Rozas et al. 2003). Monomorphic loci were necessarily excluded from the calculations of Tajima's D , F_{ST} and Fay and Wu's H . In addition, because short alignments could result in erroneous calculations of most population genetic parameters, we primarily present data from the analysis of only those loci with ≥ 100 sites available for analysis.

For demographic inference and visualization of two-dimensional SFS (2D-SFS, see [fig. 6](#)), we used the *dadi* package (Gutenkunst et al. 2009). To polarize the SNPs, we established the ancestral state for each SNP using two outgroups, *S. vernalis* and *S. gallicus*. These species' transcriptomes

were sequenced in the same way as the other samples and were aligned to the reference and SNPs scored using the same settings. Only the SNPs where both outgroups contained the same nucleotide were used in the analysis. These outgroup species are quite closely related to *S. aethnensis* and *S. chrysanthemifolius* (maximum divergence at silent sites is $\sim 5\%$), so we consider repeated mutations at the same site in both outgroups unlikely. To minimize the effects of selection on the demographic inference, only the SNPs at 4-fold degenerate sites were analyzed. Just loci with ≥ 500 sites without missing data were used ($n = 4,851$ loci). A modified version of Proseq was used to convert SNP data to the format suitable for use in the *dadi* package.

We tried several demographic scenarios implemented in *dadi* and focused on the isolation-with-migration model with population size change (*IMpre*). The parameters of the model are described in the Results, and the most likely inferred scenario for this model is shown in [figure 7](#). This model was fit to our data with all eight parameters (T_b , N_b , s , T_s , N_1 , N_2 , $M_{1 \leftarrow 2}$, and $M_{2 \leftarrow 1}$; [fig. 7](#)) estimated simultaneously, as well as in runs with the migration parameters fixed at zero. Likelihood ratio tests were used to determine whether nonzero migration parameters improved model fit to data. Repeated runs with perturbed starting parameters (*perturb_params* function in *dadi*) were used to ensure we found the global maximum. To evaluate the robustness of the parameter estimates, we used the *sample* method to create 300 bootstrap replicates of the observed SFS. The model was fitted to each of these replicates and the parameters were estimated ([fig. 7](#)).

To identify the loci putatively involved in species differentiation, we used BayeScan (Fischer et al. 2011; Foll and Gaggiotti 2008) with loci for which sequences from all 20 individuals and ≥ 100 gap-free alignment positions were present. Frequencies of each haplotype per locus were calculated from Proseq outputs and the program run with 150,000 iterations, discarding the first 50,000 as burn in. Outlier loci based on a 1% and 5% FDR were identified.

Identifying Candidate Genes Based on Expression

CLC-GW was used to calculate RPKM (Mortazavi et al. 2008) for each contig in each *S. aethnensis* and *S. chrysanthemifolius* individual as a measure of gene expression. This metric is standardized by length of the contig across all individuals and by total read number for that individual. Values were transformed by adding 0.0001 to each score (such that transcripts undetected in some individuals could still be statistically tested for expression differences) and quantile normalized. Significant differences in expression between the two species were identified using *t*-tests with an FDR of 0.05 in CLC-GW. In addition, RPKM values were analyzed in NOISeq (Tarazona et al. 2011), a nonparametric approach to identify expression differences, using default parameters. NOISeq takes into account the noise distribution of the actual data (not just the locus in question) and has been shown to be effective in controlling the number of false positives detected (Tarazona et al. 2011). We compared the list of expression candidates with those identified in a previous

microarray study of ~3,000 transcripts (Hegarty et al. 2009). Of the 224 differentially expressed probes identified by Hegarty et al. (2009), 210 were recovered from Genbank (www.ncbi.nlm.nih.gov/genbank/, last accessed October 9, 2013) and of these 128 were also present in our investigation, corresponding to 100 different contigs (i.e., some of the Hegarty et al. clones are different portions of one transcript according to our assembly). We also PCR amplified and Sanger-sequenced (using established protocols, e.g., Chapman et al. 2008; primers are listed in [supplementary table S3, Supplementary Material](#) online) a subset of the expression candidates from genomic DNA from the same individuals for a downstream analysis of F_{ST} in this subset of loci (see Results).

Analysis of Candidate Genes

Having identified candidate loci using the above tests, we carried out Fisher's exact tests in Blast2GO to determine whether certain GO terms were overrepresented, relative to the balance of the transcriptome, and hence whether the signature of selection was more pronounced across loci that share pathways or functions in common. Where only a subset of the loci had been included in the test, this subset was used as the reference set in the GO analysis. Separate tests were carried out for two groups of outliers: 1) BayeScan candidates and 2) expression candidates (outliers from the t -test and NOISeq analyses combined).

ML-HKA (Hudson et al. 1987) tests (Wright and Charlesworth 2004) were carried out in Proseq for various subsets of candidate loci, to try and identify if the pattern of polymorphism was indicative of selection at these loci and in which species there was evidence for selection. To do this, the sequence of each candidate locus from one species at a time was compared with 20 randomly chosen putatively neutral loci, using the *S. vernalis* transcriptome to calculate inter-specific divergence. A singleton SNP was added randomly to monomorphic loci because ML-HKA cannot run if $\theta = 0$. A strictly neutral model was run and then compared to a model in which the candidate was deemed under selection (100,000 iterations each). Twice the difference in the likelihoods of the two models was calculated and tested for significance using a χ^2 distribution with one degree of freedom (Wright and Charlesworth 2004).

Supplementary Material

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

Acknowledgments

The authors thank Owen Osborne for technical assistance and suggestions throughout the completion of this project, and Adrian Brennan for collecting the seed used in this paper. The authors thank WTCHG Oxford genomic facilities for generating high-throughput sequence data used in our analyses. The work was supported by the Natural Environment

Research Council (grant number NE/G017646/1 to D.A.F. and NE/G018448/1 to S.J.H.).

References

- Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, Tabin CJ. 2006. The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* 442:563–567.
- Aiuppa A, Bellomo S, Brusca L, D'Alessandro W, Di Paola R, Longo M. 2006. Major-ion bulk deposition around an active volcano (Mt. Etna, Italy). *Bull Volcanol.* 68:255–265.
- Anderson E. 1949. Introgressive hybridization. New York: John Wiley.
- Anderson E, Stebbins GL Jr. 1954. Hybridization as an evolutionary stimulus. *Evolution* 8:378–388.
- Arnold ML. 1997. Natural hybridization and evolution. Oxford: Oxford University Press.
- Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol.* 13:969–980.
- Brennan AC, Bridle JR, Wang A-L, Hiscock SJ, Abbott RJ. 2009. Adaptation and selection in the *Senecio* (Asteraceae) hybrid zone on Mount Etna, Sicily. *New Phytol.* 183:702–717.
- Brennan AC, Harris SA, Hiscock SJ. 2013. The population genetics of sporophytic self-incompatibility in three hybridizing *Senecio* (Asteraceae) species with contrasting population histories. *Evolution* 67:1347–1367.
- Bull V, Beltran M, Jiggins CD, McMillan WO, Bermingham E, Mallet J. 2006. Polyphyly and gene flow between non-sibling *Heliconius* species. *BMC Biol.* 4:11.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.
- Carels N, Bernardi G. 2000. Two classes of genes in plants. *Genetics* 154: 1819–1825.
- Chapman MA, Forbes DG, Abbott RJ. 2005. Pollen competition among two species of *Senecio* (Asteraceae) that form a hybrid zone on Mt. Etna, Sicily. *Am J Bot.* 92:730–735.
- Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, Burke JM. 2008. A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *Plant Cell* 20:2931–2945.
- Chibalina MV, Filatov DA. 2011. Plant Y chromosome degeneration is retarded by haploid purifying selection. *Curr Biol.* 21:1475–1479.
- Comes HP, Abbott RJ. 2001. Molecular phylogeography, reticulation and lineage sorting in Mediterranean *Senecio* sect. *Senecio* (Asteraceae). *Evolution* 55:1943–1962.
- Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676.
- Crisp P. 1972. Cytotaxonomic studies in the Section *Annuif* of *Senecio* [Ph.D. thesis]. London (UK): University of London.
- Crowe JH, Hoekstra FA, Crowe LM. 1992. Anhydrobiosis. *Annu Rev Plant Biol.* 54:579–599.
- Darwin C. 1859. On the origin of species by means of natural selection. London: John Murray.
- Earl DA, von Holdt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour.* 4:359–361.
- Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol.* 14:2611–2620.
- Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.
- Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
- Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends Genet.* 28:342–350.
- Filatov DA. 2009. Processing and population genetic analysis of multigenic datasets with ProSeq3 software. *Bioinformatics* 25:3189–3190.

- Fischer MC, Foll M, Excoffier L, Heckel G. 2011. Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). *Mol Ecol*. 20:1450–1462.
- Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180:977–993.
- Funk DJ. 2009. Investigating ecological speciation. In: Butlin R, Bridle J, Schluter D, editors. *Speciation and patterns of diversity*. Cambridge (UK): Cambridge University Press. pp. 195–218.
- Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* 433:481–487.
- Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. 36:3420–3435.
- Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 5:e1000695.
- Hegarty MJ, Barker GL, Brennan AC, Edwards KJ, Abbott RJ, Hiscock SJ. 2009. Extreme changes to gene expression associated with homoploid hybrid speciation. *Mol Ecol*. 18:877–889.
- Hong X, Scofield DG, Lynch M. 2006. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol*. 23:2392–2404.
- Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
- Jaillon O, Aury J-M, Noel B, et al. (56 co-authors). 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.
- James JK, Abbott RJ. 2005. Recent, allopatric, homoploid hybrid speciation: the origin of *Senecio squalidus* (Asteraceae) in the British Isles from a hybrid zone on Mount Etna, Sicily. *Evolution* 59:2533–2547.
- Kane NC, King MG, Barker MS, Raduski A, Karrenberg S, Yatabe Y, Knapp SJ, Rieseberg LH. 2009. Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution* 63:2061–2075.
- Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet*. 39:1151–1155.
- Kulathinal RJ, Stevison LS, Noor MAF. 2009. The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genet*. 5:e1000550.
- Lawnczak MKN, Emrich SJ, Holloway AK, et al. (30 co-authors). 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 330:512–514.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
- Martinsen GD, Whitham TG, Turek RJ, Keim P. 2001. Hybrid populations selectively filter gene introgression between species. *Evolution* 55:1325–1335.
- Maynard-Smith J, Haigh J. 1974. Hitch-hiking effect of a favorable gene. *Genet Res*. 23:23–35.
- Mayr E. 1942. *Systematics and the origin of species*. New York: Columbia University Press.
- Minder AM, Widmer A. 2008. A population genomic analysis of species boundaries: neutral processes, adaptive divergence and introgression between two hybridizing plant species. *Mol Ecol*. 17:1552–1563.
- Morjan CL, Rieseberg LH. 2004. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Mol Ecol*. 13:1341–1356.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 5:621–628.
- Muir G, Osborne OG, Sarasa J, Hiscock SJ, Filatov DA. 2013. Recent ecological selection on regulatory divergence is shaping clinal variation in *Senecio* on Mount Etna. *Evolution* 67:3032–3042.
- Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos Trans R Soc Lond B Biol Sci*. 367:409–421.
- Nei M. 1987. *Molecular evolutionary genetics*. New York: Columbia University Press.
- Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. 76:5269–5273.
- Nosil P. 2007. Divergent host-plant adaptation and reproductive isolation between ecotypes of *Timema cristinae*. *Am Natural*. 169:151–162.
- Nosil P. 2012. *Ecological speciation*. Oxford: Oxford University Press.
- Nosil P, Funk DJ, Ortiz-Barrientos D. 2009. Divergent selection and heterogeneous genomic divergence. *Mol Ecol*. 18:375–402.
- Nosil P, Vines TH, Funk DJ. 2005. Perspective: reproductive isolation caused by natural selection against immigrants from divergent habitats. *Evolution* 59:705–719.
- Oleksyk TK, Smith MW, O'Brien SJ. 2010. Genome-wide scans for footprints of natural selection. *Philos Trans R Soc Lond B Biol Sci*. 365:185–205.
- Osborne OG, Batstone TE, Hiscock SJ, Filatov DA. 2013. Rapid speciation with gene flow following the formation of Mount Etna, revealed by transcriptome sequencing of endemic Etnean *Senecio* species. *Genome Biol Evol*. 5:1704–1715.
- Papadopoulos AST, Baker WJ, Crayn D, Butlin RK, Kynast RG, Hutton I, Savolainen V. 2011. Speciation with gene flow on Lord Howe Island. *Proc Natl Acad Sci U S A*. 108:13188–13193.
- Pavey SA, Collin H, Nosil P, Rogers SM. 2010. The role of gene expression in ecological speciation. *Ann N Y Acad Sci*. 1206:110–129.
- Payseur BA. 2010. Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Mol Ecol Res*. 10:806–820.
- Payseur BA, Nachman MW. 2005. The genomics of speciation: investigating the molecular correlates of X chromosome introgression across the hybrid zone between *Mus domesticus* and *Mus musculus*. *Biol J Linn Soc Lond*. 84:523–534.
- Przeworski M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* 160:1179–1189.
- Renaut S, Nolte AW, Bernatchez L. 2010. Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Mol Ecol*. 19:115–131.
- Rieseberg LH. 1995. The role of hybridization in evolution: old wine in new skins. *Am J Bot*. 82:944–953.
- Rieseberg LH, Whitton J, Gardner K. 1999. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics* 152:713–727.
- Ross RIC. 2010. *Local adaptation and adaptive divergence in a hybrid species complex in Senecio* [PhD thesis]. Oxford (UK): University of Oxford.
- Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.
- Sadori L, Zanchetta G, Giardini M. 2008. Last glacial to holocene palaeoenvironmental evolution at Lago di Pergusa (Sicily, Southern Italy) as inferred by pollen, microcharcoal, and stable isotopes. *Quat Int*. 181:4–14.
- Savolainen V, Anstett MC, Lexer C, Hutton I, Clarkson JJ, Norup MV, Powell MP, Springate D, Salamin N, Baker WJ. 2006. Sympatric speciation in palms on an oceanic island. *Nature* 441:210–213.
- Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*. 78:629–644.
- Schemske DW, Bradshaw HD Jr. 1999. Pollinator preference and the evolution of floral traits in monkeyflowers (*Mimulus*). *Proc Natl Acad Sci U S A*. 96:11910–11915.
- Schluter D. 2009. Evidence for ecological speciation and its alternative. *Science* 323:737–741.
- Schwarz D, Robertson HM, Feder JL, Varala K, Hudson ME, Ragland GJ, Hahn DA, Berlocher SH. 2009. Sympatric ecological speciation meets

- pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics* 10:633.
- Scotti-Saintagne C, Mariette S, Porth I, Goicoechea PG, Barreneche T, Bodenes K, Burg K, Kremer A. 2004. Genome scanning for interspecific differentiation between two closely related oak species *Quercus robur* L. and *Q. petraea* (Matt.) Liebl. *Genetics* 168:1615–1626.
- Slatkin M. 1985. Gene flow in natural populations. *Annu Rev Ecol Syst.* 16:393–430.
- Slatkin M, Wiehe T. 1998. Genetic hitch-hiking in a subdivided population. *Genet Res.* 71:155–160.
- Smadja C, Butlin RK. 2011. A framework for comparing processes of speciation in the presence of gene flow. *Mol Ecol.* 20:2123–2140.
- Stebbins GL. 1959. The role of hybridisation in evolution. *Proc Am Philos Soc.* 103:231–251.
- Stölting KN, Nipper R, Lindtke D, Caseys C, Waeber S, Castiglione S, Lexer C. 2013. Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Mol Ecol.* 22:842–855.
- Storz JF. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol.* 14:671–688.
- Strasburg JL, Rieseberg LH. 2008. Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris*—large effective population sizes and rates of long-term gene flow. *Evolution* 62: 1936–1950.
- Strasburg JL, Sherman NA, Wright KM, Moyle LC, Willis JH, Rieseberg LH. 2012. What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philos Trans R Soc Lond B Biol Sci.* 367:364–373.
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
- Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. 2011. Differential expression in RNA-seq: a matter of depth. *Genome Res.* 21: 2213–2223.
- Vicre M, Farrant JM, Driouich A. 2004. Insights into the cellular mechanisms of desiccation tolerance among angiosperm resurrection plant species. *Plant Cell Environ.* 27:1329–1340.
- Whitehead A, Crawford DL. 2006. Variation within and among species in gene expression: raw material for evolution. *Mol Ecol.* 15: 1197–1211.
- Wolf JBW, Lindell J, Backstrom N. 2010. Speciation genetics: current status and evolving approaches. *Philos Trans R Soc Lond B Biol Sci.* 365:1717–1733.
- Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc Natl Acad Sci U S A.* 84:9054–9058.
- Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet.* 8:206–216.
- Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159.
- Wright S. 1951. The genetical structure of populations. *Ann Eugen.* 15: 323–354.
- Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* 168: 1071–1076.
- Wu C-I. 2001. The genic view of the process of speciation. *J Evol Biol.* 14: 851–865.
- You FM, Huo N, Gu YQ, Lazo GR, Dvorak J, Anderson OD. 2009. ConservedPrimers 2.0: a high-throughput pipeline for comparative genome referenced intron-flanking PCR primer design and its application in wheat SNP discovery. *BMC Bioinformatics* 10:331.