Differentially expressed genes match bill morphology and plumage despite largely undifferentiated genomes in a Holarctic songbird

NICHOLAS A. MASON* † and SCOTT A. TAYLOR*†
*Department of Ecology and Evolutionary Biology, Cornell University, 215 Tower Rd., Ithaca, NY 14853, USA, †Museum of Evolutionary Biology Program, Laboratory of Ornithology, Cornell University, 159 Sapsucker Woods Road, Ithaca, NY 14850, USA

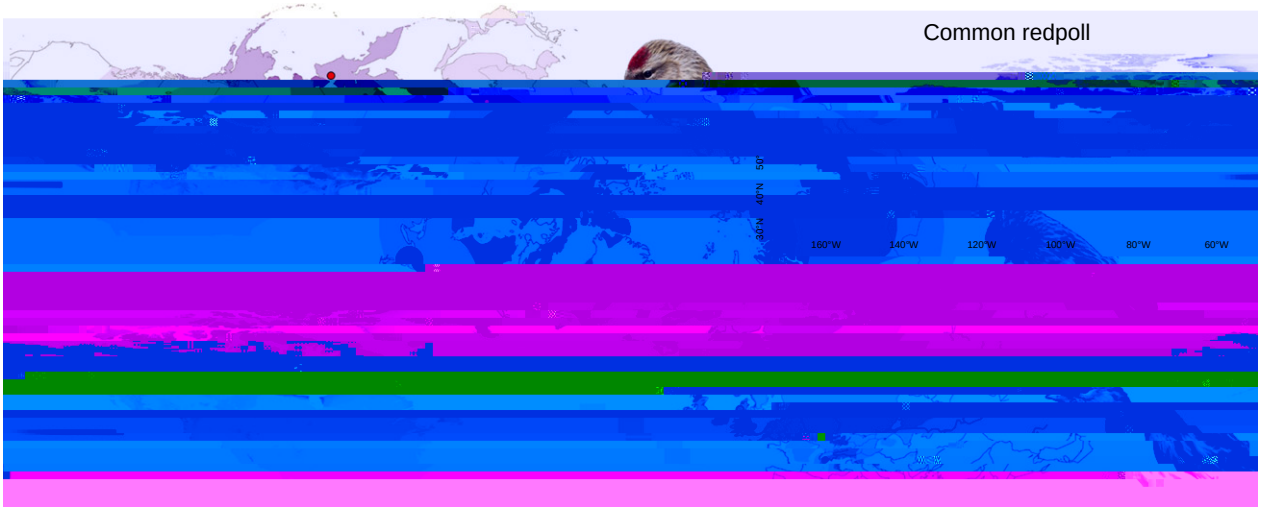
Abstract

Understanding the patterns and processes that contribute to phenotypic diversity and speciation is a central goal of evolutionary biology. Recently, high-throughput sequencing has provided unprecedented phylogenetic resolution in many lineages that have experienced rapid diversiÞcation. The Holarctic redpoll Þnches (Genus: *Acanthis*) provide an intriguing example of a recent, phenotypically diverse lineage; traditional sequencing and genotyping methods have failed to detect any genetic differences between currently recognized species, despite marked variation in plumage and morphology within the genus. We examined variation among 20 712 anonymous single nucleotide polymorphisms (SNPs) distributed throughout the redpoll genome in combination with 215 825 SNPs within the redpoll transcriptome, gene expression data and ecological niche modelling to evaluate genetic and ecological differentiation among currently recognized species. Expanding upon previous Þndings, we present evidence of (i) largely undifferentiated genomes among currently recognized species; (ii) substantial niche overlap across the North American *Acanthis* range; and (iii) a strong relationship between polygenic patterns of gene expression and continuous phenotypic variation within a sample of redpolls from North America. The patterns we report may be caused by high levels of ongoing gene ßow between polymorphic populations, incomplete lineage sorting accompanying very recent or ongoing divergence, variation in *cis*-regulatory elements, or phenotypic plasticity, but do not support a scenario of prolonged isolation and subsequent secondary contact. Together, these Þndings highlight ongoing theoretical and computational challenges presented by recent, rapid bouts of phenotypic diversiÞcation and provide new insight into the evolutionary dynamics of an intriguing, understudied non-model system.

Keywords

Lemmon 2013; McCormack et al. 2013). These novel technologies provide promising tools to study the evolution of phenotypic diversity and speciation, particularly within groups that have experienced recent and rapid diversiÞcation, which typically lack coalescence and exhibit incomplete lineage sorting (Maddison & Knowles 2006). Such Ôspecies ßocksÕ present an ongoing challenge for evolutionary biologists to discriminate true speciation events from hybrid swarms and ongoing gene ßow, especially when marked phenotypic variation is present.

Reduced-representation approaches, such as double-digest restriction-associated DNA sequencing (ddRAD-Seq) and genotyping by sequencing (GBS), are outperforming traditional Sanger-sequencing methods in their ability to help us understand the evolution of phenotypic variation in young lineages. Hybridization and incomplete lineage sorting are often common in such lineages, which can obscure evolutionary relationships (e.g. Lake Victoria cichlids, Wagner et al. 2012; Nicaraguan crater lake cichlids, Elmer et al. 2014; Heliconius butterßies, Nadeau et al. 2013; Xiphophorus Þshes, Cui et al. 2013; Jones et al.



between continents (e.g. Questiau et al. 1998; Drovetski et al. 2004, 2009).

The paucity of genetic differentiation within the redpoll complex, despite marked phenotypic variation across a Holarctic distribution, could be the result of multiple evolutionary scenarios (Marthinsen et al. 2008): redpolls may be comprised of (i) a single, undifferentiated gene pool that exhibits phenotypic polymorphism, in which phenotypic differences reßect locally adapted demes or neutral phenotypic variation within a single metapopulation; (ii) multiple gene pools that have recently diverged, in which incomplete lineage sorting has hindered the capacity of previous studies to differentiate populations or species; or (iii) multiple divergent gene pools that are actively exchanging genes through hybridization and introgression via secondary contact.

In this study, we implement high-throughput sequencing to evaluate these hypotheses by examining genome-wide variation in anonymous loci among redpolls sampled from different regions of the Holarctic. We also assess variation among transcriptome sequence data and gene expression in a subset of North American redpolls that span the phenotypic continuum described above. Finally, we use breeding season occurrence records to generate ecological niche models (ENMs) that characterize differences in suitable abiotic conditions between North American *A. ßammea* and *A. hornemanni*

Materials and methods

Sample collections and phenotyping

Molecular analyses were based on 77 individuals, including representatives of all three redpoll species currently recognized by most authorities (e.g. Clements et al. 2014), which were from different regions of their current distribution (Fig. 1, Table S1, Supporting information). Based on recently published phylogenies of the family Fringillidae (Zuccon et al. 2012), we included two white-winged crossbill (*Loxia leucoptera*) individuals as an out-group in our analyses. Because the main goal of this study was to assess genetic differentiation between redpolls with different plumage and morphology characteristics (i.e. putative species), our geographic sampling was not exhaustive from a phylogeographic perspective and we did not include representatives from all currently recognized redpoll subspecies. For this component of our study, we relied on the classiÞcations of collectors and museum curators to assign individuals to one of the three currently recognized species.

We collected 10 of the 77 redpolls included in this study on the same day at the same wintering locality (Cortland, Cortland County, NY, USA; 42.6 °N, 76.2°W; nine males and one female). These individuals were collected because they represented the broadest phenotypic variation possible within the wintering ßock. The

ßock remained at this location for over 3 weeks before collection. Therefore, individuals experienced similar environmental conditions and foraging opportunities that approximate a common garden setting. This shared experience should have reduced environmentally induced differences in gene expression between ßock members; however, we cannot completely rule out differences in microclimate or diet.

Rather than binning these individuals into putative species based on plumage characteristics and bill shape, which are known to vary continuously (Troy 1985), we measured multiple morphological characters for each individual (Table S2, Supporting information). We quantiÞed the amount of streaking on the undertail coverts and rump of each individual by taking digital photographs that were subsequently measured with IMAGEJ 1.48v (Abramoff et al. 2004). We took four measurements of beak shape (width, depth, culmen length, mandible length) for each individual using digital callipers. Bill and plumage measurements were then incorporated into a principal components analysis (PCA) to obtain multivariate dimensions of phenotypic variation (see Fig. S1, Supporting information for loadings, and Fig. S2, Supporting information for PCA scores). PCA scores were then used to assess statistical associations between phenotypic variation and multiple indices of genetic variation.

In addition to collecting genomic DNA from these 10 individuals, we also preserved separate samples of whole brain, liver and muscle in RNAlater within 25 min post-mortem for RNA-Seq data generation and gene expression analyses. Specimens were processed in the order in which they were caught, meaning that some individuals were held captive longer than others before collecting tissues. Genomic DNA and RNA samples were subsequently stored at 80 °C until library preparation.

ddRAD-Seq library preparation and sequencing

We extracted genomic DNA from each sample using Qiagen DNeasy kits (tissue protocol; Qiagen, Valencia, CA, USA), eluted the DNA in water, concentrated each sample using a vacuum centrifuge and determined the Þnal concentration of each extraction using Qubit Fluorometric Calibration (QFC; Invitrogen, Carlsbad, CA, USA). DNA extractions are archived at the Cornell Lab of Ornithology (Ithaca, NY, USA).

We prepared ddRAD-Seq libraries using a modiÞed version of the protocol outlined in Peterson et al. (2012). Following a standardizing dilution (all genomic DNA ~30 ng/ l L), we plated the samples and digested each with the restriction enzymes SbfI and MspI while ligating P1 (barcode) and P2 adaptor primers using 19 unique barcodes for each of four subsequent index groups (a total of 76 unique identiÞers—the DNA from one sample was excluded due to low quality). Each digestion reaction contained 300 ng genomic DNA, 3 l L 109

Phred score of 10. We also trimmed out any reads for which 5% of bases had a Phred score below 20. We then separated multiplexed libraries using the process_radtags function from the STACKS pipeline (Catchen et al. 2013). The Þnal Þltered, trimmed and demultiplexed data set contained 365 000 000 reads.

We pooled reads from all redpolls to perform de novo locus assembly for redpolls only using the denovo_map.pl script, which executes ustacks, cstacks and sstacks in succession and comes bundled with STACKS (Catchen et al. 2013). In brief, STACKS groups identical reads based on sequence similarity to form ÔstacksÕ, which can then be combined to form putative loci. We required a minimum of Þve reads for stack depth (-m), allowed Þve SNPs between any two stacks at a locus (-M) and Þve SNPs between any two loci when building catalogues (-n). These parameter settings performed well in a comparison of library assembly pipelines (Mastretta-Yanes et al. 2015). We allowed 20% missing data for each locus and extracted one locus per SNP using the –write_single_snp ßag when running the populations program within STACKS. One individual had to be dropped from the ddRAD-Seq assembly pipeline due to poor coverage; therefore, the Þnalized ddRAD set included 76 individuals, including 9 of 10 individuals that comprise the RNA-Seq portion of this study.

Crossbill raw sequencing reads were processed in the same manner as redpolls. The Þnal Þltered, trimmed and demultiplexed crossbill data set contained 151 511 reads. We pooled all reads from both crossbills and redpolls to perform de novo crossbill and redpoll locus assembly using the denovo_map.pl script and used the same STACKS settings detailed above.

We also identiÞed a separate panel of SNPs from the de novo transcriptome and 10 individual RNA-Seq libraries. We generated an index from our transcriptome and aligned each individual library to the reference using BWA under default settings (Li & Durbin 2009). We called SNPs from indexed alignments using the UniÞedGenotyper tool within Genome Analysis Toolkit under default settings (GATK; DePristo et al. 2011). As part of the SNP calling process, we Þltered out sites with Phred quality scores <30 and Þltered by mean depth, allele frequency and call rate and applied the BadCigarFilter using VCFTOOLS (version 3.0; Danecek et al. 2011), which removes malformed reads that start with spurious deletions. We retained a total of 215 825 out of 784 141 possible SNPs after Þltering.

Population genetic analyses

We used the Bayesian clustering program STRUCTURE v 2.3.4 (Pritchard et al. 2000) to evaluate genetic differentiation among redpolls. We analysed two sets of ddRAD-Seq SNPs using the same analytical pipeline: (i) loci assembled from only redpoll data and (ii) loci assembled with data from redpolls and the crossbill out-group. For the redpoll data set, we ran three replicate analyses for 10 000 generations following 10 000 generations of burn-in, using the ÔadmixtureÕ model across a range of K values from 1 to 5 (three replicates each), which were then averaged for population assignment scores. Because there were three putative species in this analysis, we paid speciÞc attention to results from runs where the a priori constraint on the number of population clusters was $K = 3$ (redpolls only). Results from STRUCTURE were analysed using the Evanno et al. (2005) method in STRUCTURE HARVESTER (Earl & vonHoldt 2011). For the redpoll + crossbill data set, we ran STRUCTURE as above with an a priori constraint on the number of population clusters $K = 4$ (redpolls + crossbills). We did this to ensure that our SNP data could differentiate redpolls from the out-group taxon.

To corroborate our Bayesian clustering analyses, we performed principal component analyses (PCA) on the same two sets of loci using ADEGENET v1.4 (Jombart 2008; Jombart & Ahmed 2011) and performed an analysis of molecular variance (AMOVA; ExcofÞer et al. 1992) using PEGAS

doreplicated ENMs based on 50 pooled, randomized occurrence data from both species (Graham et al. 2004; Warren et al. 2008). This procedure tests niche conservatism in the strictest sense and determines whether the

differentiation among currently recognized species or geographically isolated populations (Fig. S5A, Supporting information); the genetic PCA (Fig. S5B, Supporting information) also separated the out-group from redpolls but did not appreciably separate currently recognized redpoll species. Similarly, a PCA of the 215 825 locus data set generated from our RNA-Seq experiment (10 individuals that spanned the phenotypic continuum — see Table S2, Supporting information) did not differentiate currently described species (Fig. S6, Supporting information).

An AMOVA of the ddRAD-Seq SNP panel revealed that the overwhelming majority of genetic variation is partitioned within species (98.11%; Table 1) rather than

Supporting information). Yet, our background similarity test indicated that the abiotic conditions that characterize the distribution of A. hornemanniare more similar to those of A. ßammeathan would be expected based on the availability of habitat in North America (D = 0.579, 95% CI = 0.371–0.402; I = 0.824, 95% CI= 0.661–0.690; Fig. S10, Supporting information). Similarly, the observed similarity indices were higher than the conÞdence interval of the null distribution constructed using actual A. ßammeaoccurrence data and randomly generated A. hornemanni data (D = 0.579, 95% CI= 0.502–0.558;I = 0.824, 95% CI= 0.773–0.813; Fig. S10, Supporting information). Thus, while the abiotic niches of A. ßammeaand A. hornemanniare not completely identical, they are more similar than comparisons of either speciesÕ ENMs with null models generated from random background points throughout North America.

Discussion

Our Þndings expand upon previous studies of the evolutionary dynamics of redpolls that have relied on traditional molecular markers, such as mtDNA and microsatellites (Marten & Johnson 1986; Seutin et al. 1995; Ottvall et al. 2002; Marthinsen et al. 2008). Genome-wide panels ofary 95%Discussion

are more similar than expected based on available conditions in North America. This pattern may be present due to the dichotomous nature of the current species classiÞcation scheme, which does not account for the continuous nature of phenotypic diversity in redpolls and could underemphasize perceived differences in abiotic conditions between taxa. Individuals with intermediate plumage are placed into one of the two species categories, which may inßuence how occurrence

metapopulations exist within the genus Acanthis. We also demonstrate that individual redpolls classiÞed as different species span a phenotypic continuum, rather than discrete classes, which has been shown by previous studies (Troy 1985). Certain authorities, such as BirdLife International, already treat redpolls as a single species, and previous studies have arrived at similar conclusions (Troy 1985).

Redpolls are primarily granivorous and rely on patchily distributed tree seed crops that can vary substantially in abundance from year to year (Clement 2010a,b). As a result, redpolls are highly nomadic during the nonbreeding season and travel great distances

strongly associated with phenotypic variation included multiple genes involved in the Wnt signalling pathway (e.g. tsukushin and frizzled-3; Table S5, Supporting information). Expression levels of multiple genes involved in Wnt signalling appear to play a role in developing different bill morphologies among birds (Brugmann et al. 2010) and facial developmental pathways of various vertebrates (Brugmann et al. 2007). Additionally, the Wnt signalling pathway regulates Bmp

Greenberg R, Cadena V, Danner RM, Tattersall G (2012) Heat loss may explain bill size differences between birds occupying different habitats. PLoS ONE, 7, e40933.
Grummer JA, Bryson RW, Reeder TW (2014) Species delimitation using bayes factors: simulations and application to the sceloporus scalaris species group (Squamata: Phrynosomatidae). Systematic Biology, 63, 119–133.
Haas BJ, Papanicolaou A, Yassour M et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8, 1494–1512.
Harris MP, Norman FJ, McColl R (1965) A mixed population of redpolls in northern Norway. British Birds, 58, 288–294.
Harrison RG (1990) Hybrid zones: windows on evolutionary process. Oxford Surveys in Evolutionary Biology, 7, 69–128.
Harrison RG, Rand DM (1989) Mosaic hybrid zone and the nature of species boundaries. In: Speciation and its Consequences (eds Otte D, Ender JA), pp. 111–133. Sinauer Associates Inc, Sunderland, Massachusetts.
Harrison PW, Wright AE, Mank JE (2012) The evolution of gene expression and the transcriptome–phenotype relationship. Seminars in Cell & Developmental Biology, 23, 222–229.
Harvey MG, Brumfield RT (2014) Genomic variation in a widespread Neotropical bird (Xenops minutus) reveals divergence, population expansion, and gene flow.

at the distinctness of coastal California Gnatcatchers. Auk, 132, 380–388.
McCormack JE, Hird SM, Zellmer AJ, Carstens BC, BrumÞeld RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and

Warren DL, Glor RE, Turelli M (2010) ENMTools: a toolbox for