MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 136–150 doi: 10.1111/mec.12591

Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population

RAJIV C. MCCOY,*† NANDITA R. GARUD,‡ JOANNA L. KELLEY,§ CAROL L. BOGGS†¶ and DMITRI A. PETROV*

*Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305, USA, †Rocky Mountain Biological Laboratory, Crested Butte, CO 81224, USA, ‡Department of Genetics, Stanford University, Stanford, CA 94305, USA, §Center for Reproductive Biology, School of Biological Sciences, Washington State University, Pullman, WA 99164, USA, ¶Department of Biological Sciences, University of South Carolina, Columbia, SC 29208, USA

Abstract

The analysis of molecular data from natural populations has allowed researchers to answer diverse ecological questions that were previously intractable. In particular, ecologists are often interested in the demographic history of populations, information that is rarely available from historical records. Methods have been developed to infer demographic parameters from genomic data, but it is not well understood how inferred parameters compare to true population history or depend on aspects of experimental design. Here, we present and evaluate a method of SNP discovery using RNA sequencing and demographic inference using the program *δaδi*, which uses a diffusion approximation to the allele frequency spectrum to fit demographic models. We test these methods in a population of the checkerspot butterfly *Euphydryas gillettii*. This population was intentionally introduced to Gothic, Colorado in 1977 and has as experienced extreme fluctuations including bottlenecks of fewer than 25 adults, as documented by nearly annual field surveys. Using RNA sequencing of eight individuals from Colorado and eight individuals from a native population in Wyoming, we generate the first genomic resources for this system. While demographic inference is commonly used to examine ancient demography, our study demonstrates that our inexpensive, all-in-one approach to marker discovery and genotyping provides sufficient data to accurately infer the timing of a recent bottleneck. This demographic scenario is relevant for many species of conservation concern, few of which have sequenced genomes. Our results are remarkably insensitive to sample size or number of genomic markers, which has important implications for applying this method to other nonmodel systems.

Keywords: bottleneck, demography, lepidoptera, transcriptome

Received 15 July 2013; accepted 30 October 2013

Introduction

Demographic history shapes patterns of genetic variation within and between populations (Wright 1931). Recent methods take advantage of these patterns to infer past demographic events from genomic data sampled from natural populations (Beaumont 1999; Adams & Hudson 2004; Cornuet *et al.* 2008; Gutenkunst *et al.* 2009; Lohmueller *et al.* 2009; Lopes *et al.* 2009; Pool *et al.* 2010; Li & Durbin 2011; Lukić & Hey 2012). Inferences from genomic data supplement paleontological records to reveal ancient events in populations' history, including expansions, crashes, and migration events. While these approaches have proven invaluable, most methods of demographic inference have been empirically

Correspondence: Rajiv C. McCoy, Fax: 650 723 5920;
E-mail: rmccoy@stanford.edu

validated in systems where demographic history is known only by indirect means (e.g. alternative genetic methods or fossil evidence) and by comparing inferences to known parameters from simulated data sets. Meanwhile, all evolutionary simulations rely on particular simplifying assumptions (e.g. neutrality or absence of linked selection) that are often violated in nature and can potentially lead to inaccurate estimates of demographic parameters (Messer & Petrov 2013). It is therefore important to test methods on positive controls from natural systems with known demographic history to examine under what circumstances inferences are sensitive or robust to these violations. Similarly unexplored are issues of experimental design for generating the genomic data upon which these methods rely. A reference genome and other genomic resources are not available for many nonmodel species in which knowledge of demographic history may be desired. An approach that can inexpensively and universally survey genetic variation at the scale necessary for demographic inference can help reveal important aspects of population history in diverse study systems.

An introduced population of Gillette's checkerspot butterfly, *Euphydryas gillettii* (Nymphalidae), which has experienced recent and severe bottlenecks, offers an ideal system to examine whether demographic inference can be accurately applied to events occurring on an ecological timescale. This univoltine butterfly species inhabits meadows on eastern facing slopes of the northern Rocky Mountains. Adults fly during a 4-week period from June through mid-August with females laying clusters of more than 100 eggs on leaves of the larval hostplant, *Lonicera involucrata* (Williams *et al.* 1984). Eggs hatch in July through September, with prediapause larvae forming communal feeding webs. The larvae then overwinter in diapause within these webs until they emerge in May and June, experiencing high mortality during diapause. Post-diapause larvae move out of the web for feeding and pupate away from their host plants near the ground.

The *E. gillettii* native range spans from western Wyoming through Idaho and Montana into Alberta and British Columbia. In 1977 (33 years prior to sampling for this study), the species was intentionally introduced to a field site at the Rocky Mountain Biological Laboratory in Gothic, Colorado (CO) (Holdren & Ehrlich 1981) (Fig. 1A). Founder individuals were obtained from a population at Granite Creek, Wyoming (WY), which has since been extirpated (R. C. McCoy & C. L. Boggs, personal observation). The CO and WY habitats were intentionally matched as closely as possible, including an increase in elevation in CO accounting for the difference in latitude between the two sites. As a poor disperser with narrow habitat requirements, the introduced population of *E. gillettii* has been completely isolated from the native range by the arid Great Divide Basin, eliminating gene flow as a potentially confounding factor in our demographic analyses (Williams 1988; Boggs *et al.* 2006). Demographic data were recorded throughout these 34 generations with the exceptions of 1990–1997 and 1999–2001, during which the population was unlikely to have reached large numbers. The population established at the introduction site, persisting at 200 or fewer adult individuals for over a decade, including two separate observed bottlenecks of fewer than 25 adult butterflies (Fig. 1B). Over the past decade, the population experienced drastic fluctuations, with mark-release-recapture estimates ranging from 100 to nearly 10 000 adult individuals (Boggs *et al.* 2006; C. L. Boggs, unpublished).

Using this unique ecological system, our study demonstrates that multiplex cDNA sequencing (RNA-seq) can inexpensively generate sufficient polymorphism data to perform demographic inference in an ecological model species with no pre-existing genomic resources. We used the program $\delta a \delta i$ (Gutenkunst *et al.* 2009) to

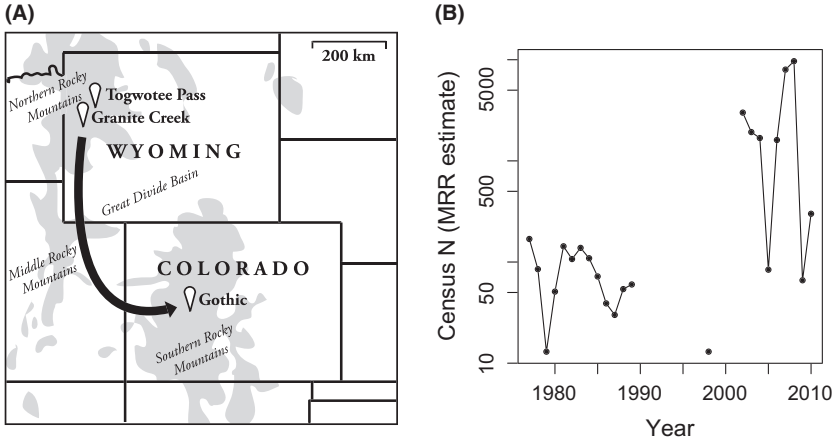

Fig. 1 Documented history of the *Euphydryas gillettii* introduction. (A) In 1977, *E. gillettii* were intentionally introduced to Rocky Mountain Biological Laboratory, Gothic, CO, from propagules obtained from Granite Creek, Wyoming (WY). Contemporary samples were obtained from Gothic as well as a site at Togwotee Pass, WY, a proxy for the now-extirpated Granite Creek source population. (B) Mark–release–recapture (MRR) estimates of adult population size in the Colorado population. The *y*-axis is depicted on a log scale to show fluctuations at very small population sizes.

infer parameters of demographic models that best fit the genomic data. The programme uses a numerical solution of a multipopulation diffusion equation to calculate the expected allele frequency spectrum for a specified demographic model, then performs optimization to find the values of the parameters which maximize the likelihood of the data given the model. This numerical approach is fast and overcomes the need for computationally demanding coalescent simulations as implemented by other approaches such as approximate Bayesian computation (ABC) (e.g. Beaumont *et al.* 2002) and Markov chain Monte Carlo (MCMC) methods (e.g. Drummond *et al.* 2012). We chose to use the δaδi software for our analyses because (i) frequency spectra can be generated from any class of polymorphic marker and the method can thereby be generalized to any large-scale genomic data set; (ii) as models are fit to the frequency spectra alone, results can be more easily interpreted as compared to more complex methods relying on many summary statistics; and (iii) the δaδi's application programming interface facilitates performance analyses to help understand how inference depends on various aspects of experimental design. Future work may compare results from different approaches to demographic inference using the same data, but such an analysis is beyond the scope of this study, which is focused on the demonstration that the frequency spectrum generated from a single data set contains sufficient information to reveal recent demographic history in a nonmodel species. The program δaδi has been widely applied, including investigation of the demographic history of humans (Gutenkunst *et al.* 2009), rice (Molina *et al.* 2011), orangutans (Locke *et al.* 2011) and other species.

Our study leverages detailed knowledge of ecology and population history of the unique *E. gillettii* system to evaluate parameter estimates and provide an important positive control in the case of recent bottlenecks, a demographic scenario that applies to many nonmodel species of conservation concern. We outline a widely applicable method for marker discovery and genotyping as well as discuss experimental considerations for studying recent bottlenecks in other nonmodel systems.

Materials and methods

Population sampling and library preparation

Eight third instar larvae were sampled from each of two field sites in September 2010: Togwotee Pass, Teton County, WY and Rocky Mountain Biological Laboratory, Gunnison County, CO. The Togwotee Pass population serves as a proxy for the now-extirpated population from Granite Creek, Teton County, WY, which is located approximately 40 km southwest of the Togwotee Pass site. The Granite Creek population, from which the CO population is derived, presumably maintained some connectivity with the Togwotee Pass population and with the rest of the *E. gillettii* metapopulation scattered throughout the Gros Ventre Wilderness. Larvae were collected and shipped alive in refrigerated containers, allowing them to clear their guts before freezing at −80 °C.

Population genomic studies encounter a common trade-off between the number of genomic markers covered at sufficient depth and the number of individuals genotyped. Faced with this trade-off, we decided to use RNA-sequencing (RNA-seq) of pooled, barcoded samples as a method to capture a reduced representation of the genome. This method allowed us to build a reference transcriptome and to discover variants from a single data set. In contrast to restriction-site-associated DNA sequencing (RAD-seq) or other methods of reduced representation, RNA-seq is biased towards discovery of variation in coding regions (Davey *et al.* 2011). By contrasting results of demographic inference using synonymous vs. nonsynonymous SNPs, we also sought to understand the impact of selection on demographic inference, which may be a confounding factor for certain experimental designs.

Total RNA was extracted from each of 16 whole larvae using a standard Trizol RNA isolation protocol. Samples were treated with the TURBO DNA-free kit (Ambion) according to manufacturer's protocol to remove DNA contamination. Samples with the highest quality (i.e. the least evidence of small RNA fragments on Bioanalyzer (Agilent) traces) were used for downstream library preparation. RNA integrity number (RIN) is not a reliable metric for this species as *E. gillettii* ribosomal RNA apparently harbours a hidden break that causes the 28S rRNA to fragment and comigrate with the 18S rRNA (Winnebeck *et al.* 2010).

To prepare cDNA libraries for the selected 16 samples, we used the TruSeq RNA Sample Preparation Kit (Illumina). This protocol includes poly-A mRNA selection, enzymatic fragmentation, first- and second-strand cDNA synthesis, end-repair, 3′ adenylation, adapter ligation and PCR amplification. Sample preparation proceeded according to the manufacturer's protocol, except for the adapter ligation step during which we incorporated custom adapters (synthesized by IDT) with 8 bp barcodes unique to each of the 16 libraries. Libraries were pooled and sequenced on a single lane of the Illumina HiSeq 2000 platform at the Stanford Center for Genomics and Personalized Medicine. Over 100 million 2 × 100 bp paired-end reads passed quality filtering and were utilized in downstream analyses.

Transcriptome assembly and annotation

We sought to assemble the *E. gillettii* transcriptome *de novo* as a reference to which to map individual sample data to discover population variation. We first demultiplexed individual sample data *in silico* according to the unique 8 bp barcodes, then trimmed these barcode sequences along with adenine overhangs (9 bp total) from the beginning of reads. We used the FastQC quality control tool <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> to evaluate the processed reads' qualities. Based on these metrics, we performed dynamic read trimming, removing ambiguous base calls at the end of FASTQ reads with the FASTX-Toolkit <http://hannonlab.cshl.edu/fastx_toolkit/index.html>. We discarded reads containing adapter and primer contamination using TagDust (Lassmann *et al.* 2009), and any remaining orphan reads were discarded.

In preparation for *de novo* transcriptome assembly, we pooled reads from all 16 libraries, then input these data to the de Bruijn graph-based assembler Trinity (Grabherr *et al.* 2011). The Inchworm module of Trinity generates a kmer catalogue and performs greedy extension based on kmer overlap. Using a range of kmer lengths during assembly can potentially improve sensitivity and allow reconstruction of transcripts with a wider range of expression levels (Schulz *et al.* 2012). We therefore modified the Trinity (version r2012-10-05) source code (C. W. Wheat, personal communication) to perform six separate assemblies with six kmer lengths (odd values from $k = 21$ to $k = 31$). We limited assembly to odd kmer lengths because even kmers may be palindromic reverse complements of themselves and introduce ambiguity to the de Bruijn graph. Assemblies were conducted on the Stanford SCG3 computing cluster with 120G of RAM. We calculated standard assembly metrics (contig number, assembly length, N50) for each of these assemblies and used blastx (Altschul *et al.* 1997) to search for homology between our contigs and a custom database of lepidopteran peptides downloaded from InsectaCentral (Papanicolaou *et al.* 2008). We assessed the degree of overlap among assemblies by comparing composition of blastx hits to the InsectaCentral lepidopteran protein database with e-value <1e-05 and alignments covering >80% of the targets' length. Based on the apparent similarity in length and content for assemblies using different kmer lengths, we selected the $k = 31$ assembly for downstream analysis to reduce the possibility that repetitive regions would produce spurious SNPs. In our case, the challenge of removing redundancy outweighed the possible gain in sensitivity of combining multiple kmer assemblies.

As quality control, we evaluated the $k = 31$ assembly based on homology to protein databases of three lepidopteran species. We first selected the longest contig sequences from each Trinity subcomponent, as multiple contigs deriving from a single subcomponent can share exons and may therefore be partially redundant. We used reciprocal blast searches to compare the *E. gillettii* transcriptome assembly with protein databases from the silkmoth (*Bombyx mori*), monarch (*Danaus plexippus*) and postman butterfly (*Heliconius melpomene*). We used blastx to search *E. gillettii* transcripts against these databases and tblastn (Altschul *et al.* 1997) to reciprocally search the protein databases against the *E. gillettii* transcriptome. Here, we report the number of unique hits with e-value <1e-03, as well as the fraction of each reference database hit by the query database (Table S2). We then limited these assessments to the small subset of contigs that harboured SNPs that we discovered downstream in our pipeline and used for demographic inference. For these contigs, we report the number of unique hits with e-value <1e-03 and the number of these hits that cover >80% of reference proteins or 50% of reference contigs (Table S3). Shorter alignment length is expected for focal species to *E. gillettii* because UTRs will not be aligned when blasting protein sequences to mRNA transcripts. We also used blastx (Altschul *et al.* 1997) to search SNP-containing contigs against the NCBI nr database, assessing the top species hits (e-value <1e-03) for all contigs as a quality control.

SNP discovery

In order to identify SNPs for the generation of site frequency spectra, each sample's preprocessed reads were mapped to the newly generated Trinity reference using BWA (version 0.6.2) (Li & Durbin 2009). We used SAMtools (version 0.1.18) to extract only uniquely aligned reads (Li *et al.* 2009). SNPs were discovered in the filtered multisample alignments using the GATK (version 2.3) UnifiedGenotyper algorithm with default parameters. We found that many called variants exhibited an extreme excess of heterozygote genotypes as well as deviation from the expected 50:50 allele balance (i.e. proportion of reads supporting the reference vs. alternative allele). In some cases, several linked variants exhibited these patterns. We suspected that these observations were due to an abundance of closely related paralogs or other repetitive sequences. In the case that one member of a paralog family is expressed at a low level, it may not be represented in the reference sequence and reads derived from this gene will map to its highly expressed, assembled paralog. Recent work supports the conclusion that a large proportion of called SNPs from RNA-seq data are indeed false positives due to hidden paralogy (Gayral *et al.* 2013).

To therefore reduce potential false positives, we modified our pipeline to allow only one mismatch per aligned read. We then used a hard filter to extract potential false SNPs with at least one sample sequenced to ≥10× coverage with reads supporting both alleles and >75% of reads supporting the reference allele. We likewise extracted putative true SNPs for which all samples were sequenced to ≥10× coverage, and any individual with nonzero counts of each allele had an allele balance between 30 and 70%. The resulting sets of 965 putative false SNPs and 6834 putative true SNPs were used to train the GATK Variant Quality Score Recalibration (VQSR) tool (Depristo *et al.* 2011) and classify all 42620 raw SNP calls as true or false at various sensitivity thresholds. The VQSR procedure, as implemented here, uses a Gaussian mixture model to distinguish true and false variants based on allele balance, the inbreeding coefficient (a measure of deviation from Hardy–Weinberg equilibrium) and mapping quality. We then extracted a final variant set consisting of SNPs that passed VQSR at a truth sensitivity threshold of 0.90 and had at least 6× coverage per sample in at least 12 of the 16 samples.

We annotated SNPs as synonymous, nonsynonymous or untranslated by identifying open reading frames (ORFs) with the program OrfPredictor (version 2.3) (Min *et al.* 2005). OrfPredictor uses homology information from blastx (to the InsectaCentral lepidopteran peptide database, in our case) as well as *de novo* prediction based on intrinsic signals in the absence of blastx results. Using ORF predictions, we translated sequences after substituting the alternative SNP, classifying variants as nonsynonymous if the substitution altered the amino acid sequence.

In order to limit the potentially confounding effects of selection on demographic inference, we first confined analyses to high-confidence synonymous SNPs discovered by our pipeline. These SNPs were used to generate a joint site frequency spectrum for input to δaδi (version 1.6.3). To incorporate information from all markers and deal with instances of missing data, we projected the frequency spectrum down to six samples (12 alleles) per population. The projection method of δaδi uses a hypergeometric distribution to effectively average over all possible results of sampling 6 alleles per population from the total number of genotype calls at each SNP (Gutenkunst *et al.* 2009).

For visualization of the genetic data used for demographic reconstruction, we generated a heatmap of the folded (i.e. unpolarized) joint frequency spectrum of all SNPs using the package *ggplot2* within the R statistical environment (Fig. 3A) (Wickham 2009; R Core Team 2013). We also performed Q-mode principal component analysis on the genotype matrix using the *ade4* package (Fig. 3B) (Dray & Dufour 2007). Genotypes were encoded as 0, 1 and 2, representing homozygous for the major allele, heterozygous and homozygous for the minor allele, respectively.

Demographic inference

For each of these three models, best-fit parameter estimates were inferred using synonymous SNPs conforming to our aforementioned filtering criteria (Table 1). We then simulated Poisson sampling from the frequency spectrum with the built-in sampling method in δaδi to generate 1000 bootstrap samples per model. Confidence intervals were constructed using empirical quantiles of the bootstrap distribution. All model parameters were positive by definition, so in cases where greater than 2.5% of bootstrap results fell at the lower boundary of the parameter space, the lower end of the confidence interval is reported as zero. We specified three simple demographic models in δaδi, the first and last of which reflect known demographic history.

Model A. Model A, a two population model (Fig. 2A), was fit using data from both the WY and CO populations. In this model, we inferred the parameters τ_{SPLIT}, η_{WY} and η_{CO}, which specify the timing of the CO population establishment (or alternatively, the bottleneck duration), the effective size of the WY population and the effective size of the CO population, respectively.

Population sizes were inferred in units relative to an ancestral effective population size arbitrarily set at one, while time was inferred in coalescent units of τ, where $\tau \cdot 2N_{ANC} = T$ generations. To therefore compare τ_{SPLIT} to the timing of the introduction known from the demographic record, we estimated the effective population size of the CO population (N_{CO}). We derived annual population estimates from mark–release–recapture estimates of census N or counts of egg clusters, as detailed in Boggs *et al.* (2006) (Table S4, Supporting information). For years during which mark–release–recapture was not performed, we used a regression model incorporating significant weather variables to estimate adult population size (Table S5, Supporting information). We accounted for deviations from 1:1 sex ratios with the equation $N_e = 4N_mN_f/(N_m + N_f)$, where N_m and N_f are the annual census estimates of adult males and females, respectively (Hedrick 2011). For years during which mark–release–recapture data were insufficient to generate separate counts of males and females, we applied the average reduction in N_e due to deviation from 1:1 sex ratio of $0.94N$. The multigeneration estimate of N_e is then the harmonic mean of these sex-ratio-corrected single-generation estimates (N_i) across t generations: $1/N_e = \frac{1}{t}\sum_{i=1}^{t}(1/N_i)$ (Hedrick 2011). We

Table 1 Best-fit parameter estimates for alternative demographic models fit to various portions of the data. Models correspond to Fig. 2. Fixed parameters are indicated in bold, and 95% confidence intervals are indicated in brackets. Effective population sizes are reported with respect to an ancestral population arbitrarily set at $\eta_{ANC} = 1$, times are reported in units of τ, where $\tau \times 2N_{ANC} = T$ generations, and migration rates in units of $M_{i\to j}$, where $M_{i\to j}/2N_{ANC} = m_{i\to j}$, the proportion of individuals in population j who are new migrants from population i every generation. Likelihoods and AIC are directly comparable for models A, B1 and B2, which represent nested models fitted with the same data

Data	Model	η_{WY}	η_{CO}	$M_{WY\to CO}$	$M_{CO\to WY}$	τ_{SPLIT}	Log likelihood	AIC
WY & CO synonymous	A	0.922 [0.673–1.253]	0.104 [0.076–0.137]	NA	NA	0.066 [0.047–0.087]	−211.86	429.72
WY & CO synonymous	B1	0.884 [0.671–1.166]	0.119 [0.082–0.171]	0.080 [0.051–0.121]	NA	0.887 [0–2.056]	−211.01	430.02
WY & CO synonymous	B2	0.893 [0.649–1.190]	0.121 [0.083–0.167]	0.081 [0.051–0.122]	0.906 [0–1.911]	0.002 *	−211.00	432.00
CO synonymous	C	NA	**0.104**	NA	NA	0.048 [0.029–0.143]	−21.65	NA
WY & CO synonymous, nonsynonymous, & UTR	A	1.320 [0.936–1.838]	0.173 [0.121–0.230]	NA	NA	0.117 [0.083–0.156]	−284.17	NA

WY, Wyoming; CO, Colorado; AIC, Akaike information criterion.
*In model B2, bootstrap estimates of τ_{SPLIT} were highly erratic and non-normally distributed, so the confidence interval is not reported.

then incorporated a literature-derived estimate of variance in reproductive success based on cage experiments in *Bicyclus anynana* (Nymphalidae), further reducing N_e to $0.60N$ (Brakefield *et al*. 2001). This reduction is consistent with data from several species within Nymphalidae that suggest that nearly half of males do not mate (Boggs 1979; Oberhauser 1989, C. L. Boggs, in preparation). Upon incorporating each of these factors, we generated a rough estimate of $N_{CO} = 34$. This estimate was used to calculate an estimate of $N_{ANC} = N_{CO}/\eta_{CO}$ and scale all inferred demographic parameters to units of individuals (for population size parameters) and generations (for time parameters).

We wish to emphasize that there are many sources of uncertainty that affect our estimate of N_{CO}, including several factors for which we did not account in interest of simplicity. Variance due to sampling of the frequency spectrum and error in the regression models are easily quantified and are reported here (Table 1, Tables S4 and S5). Countless other potential sources of error, including factors such as the effect of early male emergence (protandry), fine scale population structure and assortative mating are not quantified here. The final scaling factor should therefore be regarded as a rough estimate to demonstrate that the frequency spectrum generated from expressed SNPs contains sufficient information to perform such inference. Nevertheless, the estimate of N_{CO} is independent of the genetic data and based on intensive field survey over several decades, a rare advantage of this ecological system.

Model B. In models B1 and B2, we extended model A to infer recent migration between the WY and CO populations (Fig. 2B). Although we know that no such migration actually occurred, we were interested in inferring migration because in many systems, researchers will not have pre-existing knowledge that precludes gene flow. In these cases, inferences of gene flow may confound inference of other demographic parameters. In model B1, we inferred the rate of unidirectional migration from WY to CO (Fig. 2B1). If barriers to migration were absent, this scenario would be plausible as the native range populations could act as a source to the smaller CO sink population. In model B2, we inferred separate migration rates in each direction (Fig. 2B). In each case, inferred migration rates are reported in units of $M_{i\to j}$, where $M_{i\to j} = 2N_{ANC}m_{i\to j}$ and $m_{i\to j}$ are defined as the proportion of individuals in population j that are new migrants from population i every generation. We then performed model selection by calculating the Akaike information criterion (AIC) for each of the migration models as well as the model with no migration, preferring the model with the minimum AIC value (Akaike 1974).

Model C. Model C (Fig. 2C) was fit using data from only the CO population. Inferring demographic history from only one population allowed us to understand how the addition of data from the second (proxy ancestral) population affected precision in demographic inference. In this model, an ancestral population experiences

a bottleneck starting at time τ_{SPLIT} in the past and extending to the time of sampling. This bottleneck is modelled by a change in the effective population size from η_{ANC} to η_{CO} at time τ_{SPLIT}. Because τ_{SPLIT} and η_{CO} are confounding variables, we fixed η_{CO} to the best-fit estimate from the two-dimensional model and inferred τ_{SPLIT}.

Nonsynonymous SNPs

To test the effect of selection on parameter estimates, we repeated the demographic inference procedure for model A (Fig. 2A), this time fitting our model to the full data set of 6349 high-confidence synonymous, nonsynonymous and untranslated SNPs with genotype calls in at least six samples per population. We contrasted parameters estimated from this larger data set to those inferred from the smaller set of synonymous markers alone. To verify that any differences were not an artefact of the larger number of markers, we randomly subsampled the frequency spectrum to the same size as the synonymous data set (1881 SNPs), repeating this procedure 1000 times and estimating 95% confidence intervals.

Performance analyses

We were interested in the sensitivity of parameter estimates to the number of SNPs included in our analysis. We used the program *ms* (Hudson 2002) to sample varying numbers of SNP markers from eight individuals per population, simulating the frequency spectrum under the best-fit parameters of model A (Fig. 2A). Given that the relative genomic locations of SNP markers were unknown, we elected to simulate a single locus with a high recombination rate ($\rho = 2000 = 4N_{ref}r$, where r is the per-generation probability of recombination between the ends of the locus and N_{ref} is the effective size of the ancestral population). We then scaled the resulting frequency spectrum to a given number of segregating sites using the frequency spectrum manipulation functions in $\delta a \delta i$. We also tested a range of recombination rates, finding that they did not qualitatively alter our results. Our approach represents a balance between speed of simulation and a desire to account for additional variance in the frequency spectrum due to physical linkage among markers. An alternative to this approach would be to independently simulate unlinked loci, but our approach is conservative in that it accounts for correlations in coalescent history arising from physical linkage of among markers. We repeated the simulations 1000 times for each number of sampled markers and used $\delta a \delta i$ to infer the best-fit parameter estimates for each simulated data set. This procedure allowed us to examine how the variance in estimates as well as the proportion of nonconverging estimates changed as a function of sample size of SNP markers (Fig. 4A).

We were similarly interested in the sensitivity of parameter estimates to the number of sampled individuals per population. We again simulated model A (Fig. 2A) using *ms*, but incremented the number of sampled individuals from one to 10 per population, with the number of SNPs fixed at 1881. We repeated the simulations 1000 times under the best-fit parameters of model A (Fig. 2A) and again used $\delta a \delta i$ to infer the best-fit parameters estimates for each simulated data set. We then examined how variance in parameter estimates and the proportion of nonconverging estimates changed as a function of the number of individuals sampled per population (Fig. 4B).

Results

Transcriptome assembly, annotation and SNP discovery

Combining sequence data from all 16 individuals, we used Trinity to perform *de novo* assembly of the *Euphydryas gillettii* larval transcriptome. We performed separate assemblies using a range of kmer lengths for the first Trinity module called Inchworm. Each assembly produced greater than 50 000 subcomponents which contain one or more isoforms of putative transcripts. When selecting the longest contig per subcomponent, N50 length ranged from 812 to 1320 for different kmer choices (Table S1, Supporting information). Greater kmer lengths are better for distinguishing among short repetitive sequences, but may lead to a more convoluted de Bruijn graph. Based on our goal of variant discovery, we were less concerned with assembly contiguity than the presence of false-positive SNPs, so we selected the $k = 31$ assembly for all downstream analyses.

We compared the *E. gillettii* transcriptome to protein sequence data available from three other lepidopteran species using reciprocal blast searches. Our transcriptome assembly covered a large proportion (69.6–76.5%) of the proteomes of these related species (Table S2, Supporting information). We observed a comparable number of matches when searching these species' proteomes against the *E. gillettii* transcriptome. The lower fraction of hits to the target database reflects differences in the sizes of the transcriptome and proteome databases, divergence among orthologs, genes that are unique to individual species, as well as possible contamination or spurious transcripts within each database.

All downstream analyses were biased, however, towards transcripts that were sufficiently highly

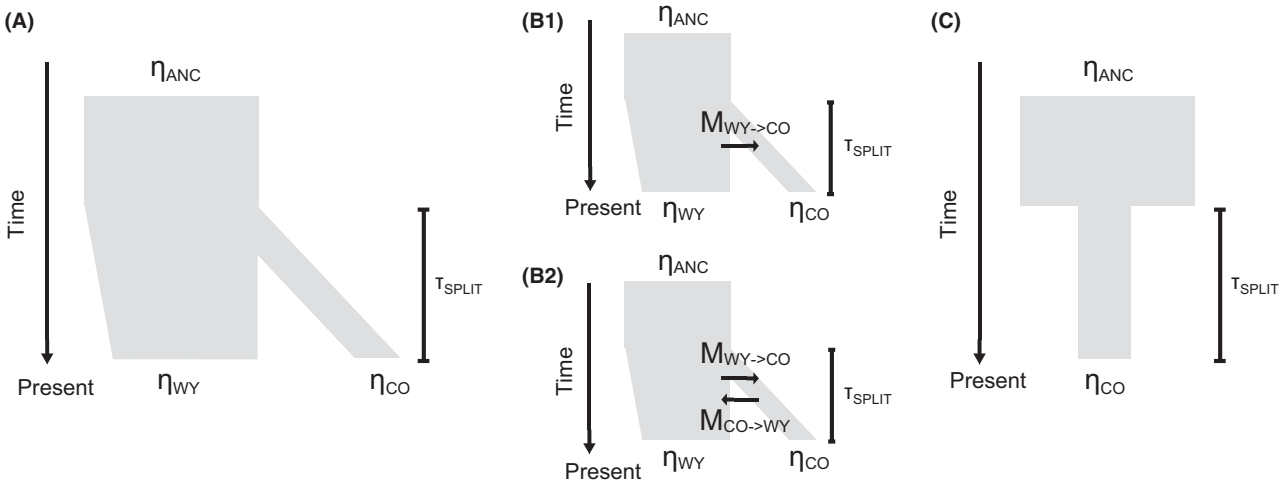

Fig. 2 Demographic models specified in *δa*δi. (A) Two-dimensional model fit with Wyoming (WY) and Colorado (CO) data. An ancestral population in WY gives rise to a derived CO population through a founding event at time τ_{SPLIT} in the past. The resulting populations in WY and CO have sizes η_{WY} and η_{CO}, respectively. (B1) Model A is extended to include possible unidirectional migration from WY to CO. (B2) Model A is extended to include possible bidirectional migration, both from WY to CO and from CO to WY. (C) Single-population model fit with CO data. An ancestral population experiences a bottleneck at time τ_{SPLIT} in the past, reducing its size to η_{CO}.

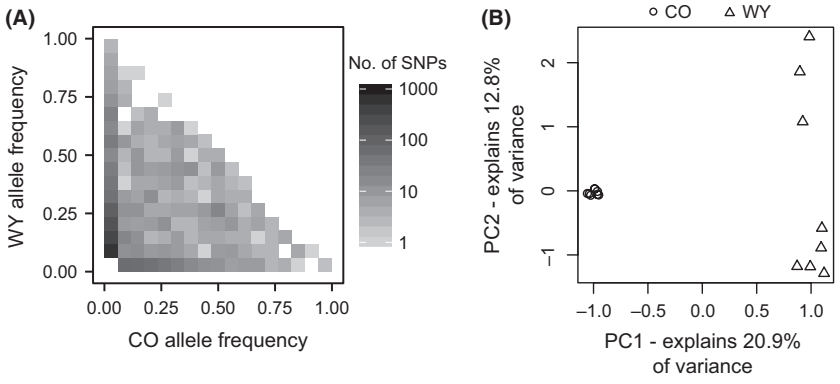

Fig. 3 Representations of the genetic data. (A) Joint allele frequency spectrum composed of all SNPs segregating in Wyoming (WY), Colorado (CO) or both populations. The frequency spectrum illustrates the loss of ancestral genetic variation in the CO population due to genetic drift during the bottleneck. Frequencies range from 0 to 16 chromosomes per population. The spectrum, displayed as a heatmap, is folded (i.e. unpolarized), as the state of the ancestral allele is unknown. (B) Individual samples plotted according to the first two principal components of the genotype matrix of all SNPs. Populations are indicated with different plotting symbols. Upon stratifying data by SNP class (synonymous, nonsynonymous, UTR), results were qualitatively similar and are not depicted. Principal component 1 separates samples according to population membership, while principal component 2 separates individuals within the WY population (within which the CO samples are nested, but tightly clustered).

expressed in enough individuals to make high-confidence genotype calls. As a result, only 2757 of the 56 536 unique contigs harboured high-confidence SNPs ultimately used for demographic inference. Higher expression levels facilitate the faithful reconstruction of mRNA transcripts, and highly expressed genes tend to be more evolutionarily conserved (Subramanian 2004). We consequently observed a higher proportion of reciprocal blast hits between this subset of *E. gillettii* transcripts and protein databases of related species (Table S3, Supporting information). Of 2408 SNP-containing contigs with significant (e-value <1e-03) hits to the NCBI nr database, 15 had top matches to plants, 99 had top matches to bacteria and only one had a top match to humans, which together represent the most likely sources of contamination in this experiment. Meanwhile, 2009 sequences had top matches to lepidopteran species. The remainder likely reflects genes that are either species-specific, highly conserved, or highly diverged and therefore do not match to lepidopteran

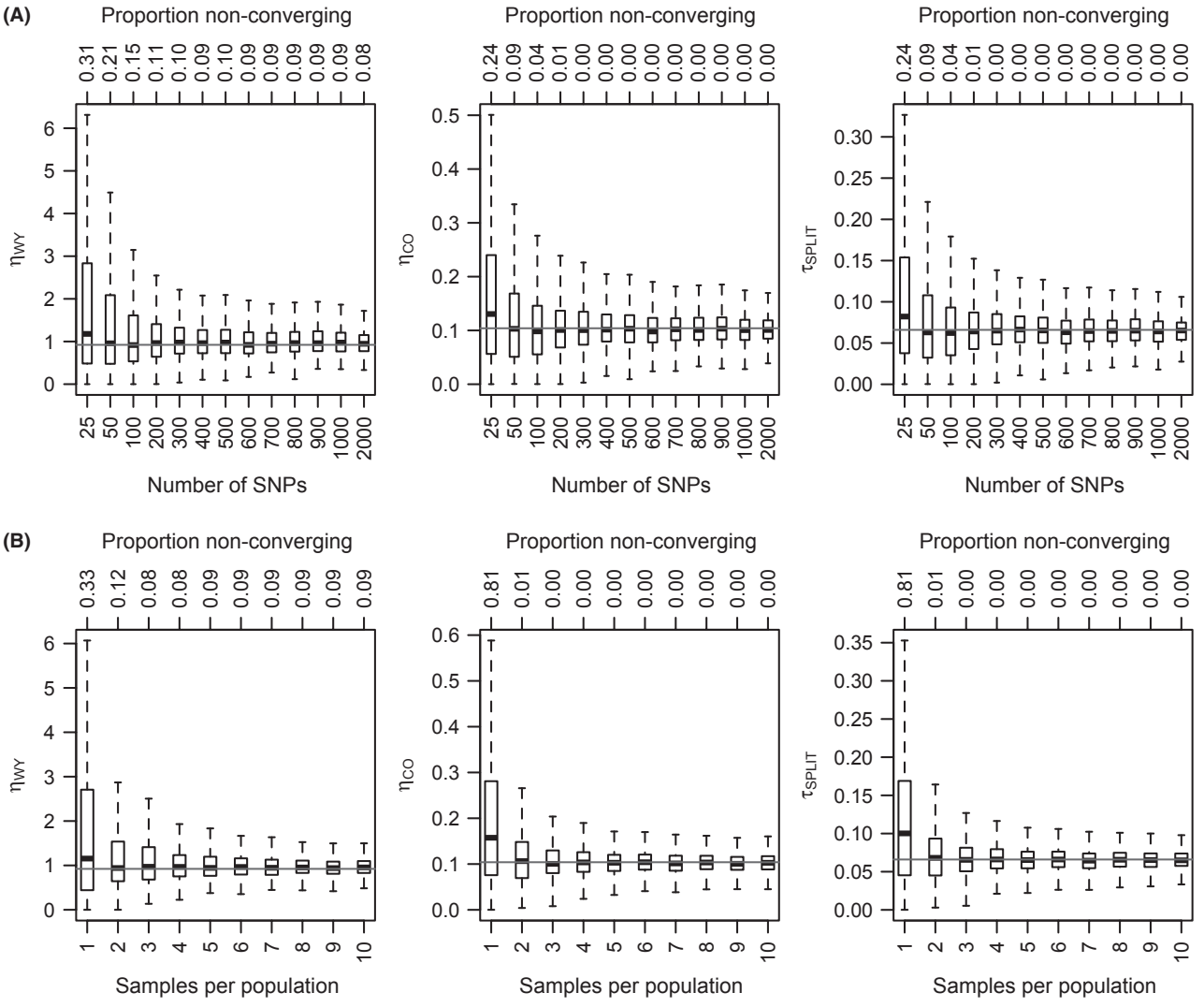

Fig. 4 Performance analysis to test the effect of number of samples and number of SNPs on demographic inference results. We removed results where estimates hit the upper or lower bounds of the set parameter range, but report the proportion of these non-converging estimates on the top axes. Simulated parameter values are indicated by the horizontal line and correspond to the best-fit estimates from model A. (A) Best-fit parameter estimates when fitting the model with varying numbers of SNPs, demonstrating that variance in estimates is relatively stable with 300 or fewer markers. Sample number is fixed at 8 per population, and simulation and model fitting are performed 1000 times for each size SNP set. (B) Best-fit parameter estimates when fitting the model with varying numbers of samples per population, demonstrating that variance in estimates is relatively stable with as few as two samples (four alleles) per population. SNP number is fixed at 1881, and simulation and model fitting are performed 1000 times for each number of samples.

reference proteins. We therefore opted against filtering SNPs based on these results, as such filtering could introduce new biases that could confound downstream demographic analyses. Together, our results suggested that spurious transcripts and contamination are rare in the portion of our assembly utilized for demographic inference.

We incorporated homology information from blast searches to all available lepidopteran protein data to identify likely ORFs using the program OrfPredictor. This allowed us to classify 2277 synonymous, 1396 nonsynonymous and 2675 UTR SNPs with at least 6× coverage per sample in at least six samples per population. As expected under purifying selection, the synonymous and nonsynonymous frequency spectra differed in shape in both the WY ($\chi^2[6,N = 1276] = 16.79$, $P = 0.010$) and CO ($\chi^2[6,N = 531] = 12.63, P = 0.049$) populations, with an excess of nonsynonymous SNPs at low frequency (WY synonymous Tajima's $D = -0.0494$, WY nonsynonymous Tajima's $D = -0.385$, CO synonymous Tajima's $D = 0.692$, CO nonsynonymous Tajima's $D = 0.505$). Of the 2277 synonymous SNPs, 1991 and

959 were segregating in the WY and CO populations, respectively. While we identified 71% of CO SNPs segregating in WY, we only identified 34% of WY SNPs segregating in CO. The asymmetry in the number and overlap of segregating sites is consistent with the founder event and subsequent bottlenecks causing substantial allelic extinction in the derived population.

Demographic inference

Model A. The demographic model, in which an ancestral population from WY splits to form the introduced CO population (Fig. 2A), reflects our knowledge of the true population history. Upon fitting this model using data from both the contemporary WY and CO populations, we recovered converging estimates of all demographic parameters (Table 1). Our model underestimated the number of low frequency SNPs that were lost in the CO population, but provided a good fit to the data overall as the model and data frequency spectra were not significantly different (χ^2[86,N = 1881.4] = 50.09,P = 0.999) (Fig. S1, Supporting information). While the effective size of the WY population (η_{WY}) was inferred to be approximately the same as the ancestral population (95% CI [0.673–1.253]), δaδi inferred a severe bottleneck (95% CI [0.076–0.137]) in the CO population (η_{CO}). These population sizes are reported with respect to an ancestral population arbitrarily set at η_{ANC} = 1. In addition, δaδi detected that the bottleneck timing (τ_{SPLIT}) was recent (95% CI [0.047–0.087]), with time reported in units of $2N_{ANC}$ generations.

Upon scaling the inferred parameters to units of individuals and generations (for population sizes and times, respectively), we found that inferred parameters were consistent with the documented history of the population. Our census-based estimate of N_{CO} = 34 placed the scaled estimate of τ_{SPLIT} based on best-fit parameters of model A between 40 and 47 generations in the past (95% CI), calculated as $2\tau_{SPLIT}(N_{CO}/\eta_{CO})$. We note, however, that this confidence interval accounts only for uncertainty in τ_{SPLIT} and η_{CO}. Uncertainty in the crude estimate of N_{CO} also contributes to uncertainty in the scaled parameter estimate, which would inflate the confidence interval beyond the reported limits. Nevertheless, our estimate of bottleneck onset is close to the known population establishment 33 generations prior to sampling, with one generation per year in this system. This result demonstrates that the joint frequency spectrum generated from RNA-seq data contains sufficient information to infer parameters of demographic scenarios occurring in the recent past.

Model B. Further analyses focused on considering the robustness of the above results to different treatments of the data and different specifications of the demographic model. First, we extended the two- dimensional demographic model to infer recent migration between the WY and CO populations (Fig. 2B), although we are confident that no such migration occurred. In many systems, however, researchers will not be able to exclude this possibility, and inferences of migration may be confounded with inferences of other demographic parameters. We therefore incorporated migration by modelling unidirectional gene flow from WY to CO ($M_{WY\rightarrow CO}$, Fig. 2B1) as well as bidirectional gene flow of potentially different magnitudes between WY and CO ($M_{WY\rightarrow CO}$ and $M_{CO\rightarrow WY}$, Fig. 2B).

For model B1, we found that δaδi inferred a low migration rate (95% CI [0.051–0.121]), but that uncertainty in the estimate of τ_{SPLIT} (95% CI [0–2.056]) dramatically increased to the point that the confidence interval included the parameter boundary of zero (Table 1). This result is not unexpected, given that migration and drift have contrasting effects on the allele frequency spectrum (Gutenkunst *et al.* 2009). To observe the same amount of drift in the joint frequency spectrum in the face of nonzero migration, bottleneck duration must be greater. However, these effects cannot be disentangled from the frequency spectrum alone, which generates uncertainty in the estimates. Estimation of τ_{SPLIT} became erratic upon adding the free parameter $M_{CO\rightarrow WY}$ in model B2, likely due to overfitting of our limited sample.

We evaluated the improvement in likelihood given the increase in model complexity by calculating the AIC for each migration model as well as the model with no migration (Akaike 1974). The model with no migration had the minimum AIC and was therefore preferred over the more complex migration models, consistent with the known demographic history of population isolation (Table 1). As models A, B1 and B2 represent nested models, we similarly applied the likelihood ratio test, finding that the model fit was not significantly improved when allowing for unidirectional (χ^2[1] = 1.70, P = 0.192) or bidirectional migration (χ^2[2] = 1.72, P = 0.423) as compared to the null model with no migration. Finally, Gutenkunst *et al.* 2009 showed that fitting data including migration with a no-migration model results in correlated residuals. Our residuals plot for model A shows no evidence of this phenomenon (Fig. S1C). In summary, our results demonstrated that while inference of migration may confound inference of other demographic parameters, model selection procedures may help indicate whether such migration actually occurred.

Model C. When we fit a simple bottleneck model to data from only the CO population (Fig. 2C), our model predictions fit the data relatively well (χ^2[3,

$N = 803.6] = 3.15, P = 0.370$). Nevertheless, bottleneck magnitude (η_{CO}) and timing (τ_{SPLIT}) have confounding effects on the site frequency spectrum and cannot be disentangled using data from a single population. We were interested, however, in the effect of the additional information from the WY population on inference of τ_{SPLIT}. We therefore fixed η_{CO} to 0.104, its best-fit estimate from the model fit using data from both populations and repeated demographic inference on the CO site frequency spectrum. With η_{CO} fixed, $\delta a \delta i$ infers a τ_{SPLIT} of 0.048. The confidence interval of τ_{SPLIT} inferred from this single-population spectrum (95% CI [0.029–0.143]) entirely includes that estimated from the joint-population spectrum in model A (95% CI [0.047–0.087]) demonstrating that we gained precision with multiple-population inference.

Nonsynonymous SNPs

We initially fit all models using only synonymous SNP data, which we presumed was important because selection can alter the frequency spectrum, confounding signatures of neutral demographic history. We examined whether this is the case for RNA-seq data by comparing inferences using only synonymous SNPs to the full data set of 6349 synonymous, nonsynonymous and UTR SNPs. In this case, best-fit estimates of η_{WY}, η_{CO} and τ_{SPLIT} significantly exceeded those inferred when fitting the model using synonymous SNP data alone (Table 1). This difference is not an artefact of the larger number of SNP markers, as randomly resampling to the same size as the synonymous data set (1881 SNPs) produced confidence intervals for η_{WY} (95% CI [0.936–1.838]), η_{CO} (95% CI [0.121–0.230]) and τ_{SPLIT} (95% CI [0.083–0.156]) that included the estimates from the full data set, but exceeded the estimates from the synonymous data alone. These results suggest that natural selection indeed distorted the frequency spectrum and changed our inferences of demographic parameters. Parameter overestimation is caused by the skew of the nonsynonymous frequency spectrum towards rare variants in both populations (Fig. S2, Supporting information). The distortion of the CO frequency spectrum for nonsynonymous SNPs is likely a carryover of purifying selection in the ancestral population, as N_{CO} was too small for selective differences to generate observable frequency differences within CO.

Performance analyses

To better understand how parameter estimates were sensitive to the number of SNP markers and the number of sampled individuals per population, we simulated frequency spectra under the best-fit parameters of demographic model A (Fig. 2A), then used $\delta a \delta i$ to infer these parameters from the simulated data. We subsampled the simulated frequency spectra for different numbers of SNP markers and different numbers of individuals. While median parameter estimates were robust even for very small marker sets (as few as 50 SNPs), variance in inferred parameters increased substantially below approximately 400 SNPs (Fig. 4A). Increasing marker sets above 400 SNPs only marginally decreased the variance in estimates and the proportion of nonconverging estimates. We likewise found that $\delta a \delta i$ performed remarkably well even with sample sizes as low as three individuals per population (Fig. 4B). Given our particular demographic scenario, sampling more than four individuals per population did not appreciably reduce variance in estimates or the proportion of nonconverging estimates.

Discussion

Our study generated the first genomic resources for Gillette's checkerspot butterfly, *Euphydryas gillettii*, using a single data set to assemble the reference transcriptome and discover genetic variation in two populations. We leveraged these population genomic data to perform demographic inference in this rare isolated system with a well-known history of recent bottlenecks. This demographic scenario is relevant to many ecological systems, including species introductions from a small number of propagules and populations of conservation concern that have experienced recent declines. We used the program $\delta a \delta i$ to accurately infer the timing of the population's introduction (and accompanying reduction in population size), providing a unique positive control given this particular demographic history. Our study complements a large body of previous work using checkerspot butterflies as model systems in conservation and metapopulation biology (Ehrlich & Hanski 2004). Within this context, this work demonstrates how genomic studies of ecological model systems can provide valuable tests of population genetic theory and methods.

SNP discovery in RNA sequence data without pre-existing genomic resources is challenging. Well-developed methods such as the GATK framework (Depristo *et al.* 2011) are designed for detecting variants in genomic DNA-derived sequence data. However, high-coverage whole-genome resequencing is currently prohibitively expensive in most eukaryotic systems, and sequence capture methods depend on *a priori* knowledge of the genome sequence to be targeted. Restriction-site-associated DNA sequencing (RAD-seq) offers one reduced representation alternative by sequencing restriction-site flanking regions in multiple individuals.

For the purpose of demographic inference, RAD-seq may in fact be preferable to RNA-seq in that highly expressed genes do not account for a large proportion of overall sequence data [although normalization methods have been devised to address this problem (Christodoulou *et al.* 2011)], purifying selection is less likely to affect these randomly dispersed genomic regions, and gene paralogy is less likely to confound marker discovery. For systems with no pre-existing genomic resources, however, researchers may desire a method that can discover neutral genomic markers for demographic inference as well as surveying functional regions. RNA-seq may be preferable in such cases because it requires no *a priori* knowledge of the genome sequence and preferentially targets transcribed regions of the genome that are more likely to be functional. As we demonstrated, this allows researchers to address not only questions about neutral effects of demographic history, but also the interplay of selection and demography in nonmodel systems. With the appropriate experimental design, the same data may also be leveraged for gene expression analysis or comparative transcriptomics between populations or between species.

Careful curation of the reference assembly, tuning of mapping parameters and stringent filtering are however necessary to extract a high-quality SNP set from RNA sequence data. Hidden paralogy generates many spurious SNP calls which can have negative effects on downstream analyses (Gayral *et al.* 2013). Here, we used heuristic SNP filtering to conservatively identify putative false positives and true positives, using these sets to train a Gaussian mixture model and classify variants. Filtering should be performed with care, as certain filtering strategies (e.g. allele frequency thresholds) could distort the resulting frequency spectrum and confound demographic inference.

We specified three basic demographic models, the first of which reflected the known demographic history and included both the WY and CO populations (Fig. 2A). We fit this model using the synonymous joint frequency spectrum, then scaled the inferred bottleneck duration (τ_{SPLIT}) based on our estimate of the effective size of the CO population. This estimate was derived from mark–release–recapture estimates of adult population size and sex ratios in *E. gillettii* (Boggs *et al.* 2006, C. L. Boggs, unpublished) as well as a correction for high variance in reproductive success as reported in other lepidopteran species (Boggs 1979; Oberhauser 1989; Brakefield *et al.* 2001). The resulting estimate of bottleneck duration of between 40 and 47 generations (95% CI) compares favourably to the documented introduction 33 generations ago. We note that the scaled values of demographic parameters carry uncertainty from both the inference procedure (due to sampling of the frequency spectrum, for which we account using the bootstrap procedure) and from uncertainty in the estimate of the scaling factor N_{CO}, for which we do not account, but discuss here. Crude methods of estimating effective population size tend to overestimate N_e, as most biological factors reduce N_e relative to census N. In particular, our consideration of how variance in reproductive success reduces N_{CO} likely underestimates the true reduction because variance in survival among egg clusters from individual females would introduce additional variability among parents. Likewise, *E. gillettii*, like many checkerspots, is highly sedentary, and population structure could further reduce N_e relative to census N (Williams 1988; Boggs *et al.* 2006). It is also likely that there is additional error in $\delta a \delta i$'s estimate due to complex evolutionary forces including genetic hitchhiking distorting the frequency spectrum relative to assumed neutrality. Nevertheless, the fact that we recover estimates of demographic parameters consistent with known demographic history suggests that these assumptions are not consequential for demographic inference, at least in this particular case.

In many cases, researchers will not have pre-existing knowledge of demographic history, yet will be interested in absolute estimates of demographic parameters rather than coalescent units relative to N_e. In such cases, estimates of N_e are often obtained from the population genetic data by estimating the parameter $\theta = 4N_e\mu$ (Watterson, 1975) and using literature-derived estimates of the mutation rate μ. Many estimates of θ, however, make the assumption of stable demographic history and can be strongly biased under certain demographic scenarios, including bottlenecks. A better approach involves inferring θ under a specified demographic model, as implemented by $\delta a \delta i$ and other methods. Mutation rate can also be estimated by performing sequence alignment between the study species and a closely related species: $\mu = D/2T$, where D is the pairwise sequence divergence and T is the divergence time in units of generations.

Our study highlights the importance of fitting multiple demographic models to test diverse demographic scenarios. We found that the model likelihood was not significantly improved by the addition of migration parameters when we extended the population split model to incorporate possible migration between WY and CO (Fig. 2B). In all other cases, however, our reported model likelihoods were not directly comparable because they were fitted with different data sets. We evaluated our models with χ^2 goodness-of-fit tests, examining whether the frequency spectrum predicted under our optimized demographic models were significantly different than the data frequency spectrum. In each case, we failed to reject the demographic models

fit with synonymous data, suggesting that these demographic models captured important aspects of the true demographic history.

In the third demographic model, we performed demographic inference using only the CO frequency spectrum, finding that uncertainty in estimates of demographic parameters was significantly greater than when including data from the WY population. Because of correlated effects on the allele frequency spectrum, bottleneck magnitude and duration could not be disentangled from these data. The addition of the WY data set (described above) added sufficient information to simultaneously infer these parameters. Upon fixing η_{CO} at its optimized value from two-dimensional demographic inference, we estimated τ_{SPLIT} similar to the two-dimensional inference. Uncertainty in the estimate increased, however, demonstrating that the addition of data from the proxy ancestral population improved precision. This result is not unexpected, as the joint frequency spectrum contains dramatically more information than the frequency spectrum of individual populations. For example, analysis of the joint-frequency spectrum revealed that of 984 total SNPs (discovered in either population and successfully genotyped in all 16 individuals), 866 were segregating in WY while only 392 were segregating in CO. Without addition of the WY data, the zero frequency class would be excluded, and inference would be limited to the CO frequency spectrum comprised of fewer markers. The joint-frequency spectrum likewise contains information about the magnitude of genetic drift by capturing the change in allele frequencies since the populations' divergence.

By contrasting inferences using synonymous data with inferences using the entire joint-frequency spectrum of synonymous, nonsynonymous and UTR SNPs, we show that natural selection distorts the frequency spectrum and leads to inaccurate parameter estimates. The fact that δaδi overestimates parameters upon inclusion of nonsynonymous and UTR SNPs is consistent with the skew of these markers towards rare variants compared with synonymous SNPs (Fig. S2). Signal in the frequency spectrum is thereby confounded because the excess of rare variants is a signature of population expansion, but also purifying selection. We should note that excluding nonsynonymous and untranslated SNPs from our analysis would not entirely resolve this issue, as purifying selection on synonymous sites as well draft due to background and/or positive selection would be reflected in the synonymous frequency spectrum. One alternative to selecting only putative neutral sites is to specify the distribution of selective effects and incorporate purifying selection in the demographic model itself (Gutenkunst *et al.* 2009) (although see Messer & Petrov (2013) for how this approach does not resolve the issue in the case of linked selection).

We demonstrate that inferences of demographic parameters are remarkably robust to both sample number and number of genetic markers. However, our results are particular to the demographic history of this system. For other systems with different demographic histories, simulations like those that we present can be useful during the planning stages of an experiment. By simulating frequency spectra for a range of demographic scenarios, researchers can evaluate the necessary number of samples and markers to achieve a given level of confidence in parameter estimates.

The δaδi approach is one of many approaches for reconstructing demographic history using population genomic data. We tested this approach on our data set as it fits demographic models using the frequency spectrum alone, which simplifies the interpretation of inference results. The flexibility of the software also facilitates various performance analyses. We remain agnostic, however, to the question of whether alternative methods would give consistent or potentially superior results. For example, Gutenkunst *et al.* (2009) point out that diffusion approximation assumes that N is large and that frequency changes are small per generation, an assumption that may be violated by an extreme bottleneck. The fact that we recover parameter estimates consistent with known demographic history, however, suggests that the approach is robust to this assumption in this particular case. The demographic history of our study population may be particularly easy to resolve due to dramatic effects on the allele frequency spectrum, whereas for other demographic scenarios that require information about the distribution of rare alleles, larger sample sizes will be required. Alternative approaches may be more appropriate for inferring parameters of different demographic scenarios on different timescales. For larger populations, the frequency spectrum contains information about substantially older events, allowing reconstruction of events occurring hundreds to thousands of generations in the past (e.g. Molina *et al.* 2011). Extreme bottlenecks introduce noise to the frequency spectrum, erasing signatures of ancient events.

Positive controls are important for understanding the circumstances under which demographic inference from genomic data is sensitive to unrealistic model assumptions and simplifications as well as particular methods of data generation. Our study provides one such positive control in a particularly well-studied system, demonstrating that it is possible to recover estimates of demographic parameters numerically consistent with known demographic history. The ability to recover information about past bottlenecks from patterns in genetic data is particularly important because bottlenecks increase the risk of population extinction (Whitlock 2000). Our RNA-seq-based approach provides

a means to simultaneously perform marker discover and multisample genotyping in systems with no existing genomic resources. We advocate more positive controls in diverse ecological model systems, leveraging detailed knowledge of species' life history for demographic modelling. Meanwhile, application of this multiplex RNA-seq approach in nonmodel species permits the study of transcribed gene sequence and expression levels while also generating polymorphism data to accurately infer recent bottlenecks. Together, these analyses from genomic data can elucidate important aspects of species' ecology and conservation status.

Acknowledgements

The authors would like to thank Jamie Walters and Diamantis Sellis for initial discussions during the conception of this experiment. Thanks also go to Alan Bergland for advice for library preparation and initial data analyses. We thank Sarah Lummis for collecting Wyoming samples. Thank you also to the many undergraduate field assistants involved in mark–release–recapture studies at Rocky Mountain Biological Laboratory. N.R.G. is supported by the NSF GRFP. The work was supported by the NIH Grants RO1GM100366 and RO1GM097415 to DAP.

References

Adams AM, Hudson RR (2004) Maximum-likelihood estimation of demographic parameters using the fre quency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, **168**, 1699–1712.

Akaike H (1974) A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**, 716–723.

Altschul SF, Madden TL, Schäffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Beaumont MA (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Boggs CL (1979) Resource allocation and reproductive strategies in several heliconiine butterfly species. Ph.D. thesis, University of Texas at Austin.

Boggs CL, Holdren CE, Kulahci IG, *et al.* (2006) Delayed population explosion of 588 an introduced butterfly. *Journal of Animal Ecology*, **75**, 466–475.

Brakefield PM, El Filali E, Van der Laan R, Breuker CJ, Saccheri IJ, Zwaan B (2001) Effective population size, reproductive success and sperm precedence in the butterfly, Bicyclus anynana, in captivity. *Journal of Evolutionary Biology*, **14**, 148–156.

Christodoulou DC, Gorham JM, Herman DS, Seidman JG (2011) Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Current Protocols in Molecular Biology*, **94**, 4.12.1–4.12.11.

Cornuet JM, Santos F, Beaumont MA, *et al.* (2008) Inferring population history with DIY ABC: a user friendly approach to approximate Bayesian computation. *Bioinformatics*, **24**, 2713–2719.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Depristo MA, Banks E, Poplin R, *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Publishing Group*, **43**, 491–498.

Dray S, Dufour A (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22**, 1–20.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.

Ehrlich PR, Hanski I (2004) *On the Wings of Checkerspots. A Model System for Population Biology*. Oxford University Press, New York, NY, USA.

Gayral P, Melo-Ferreira J, Glémin S, *et al.* (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genetics*, **9**, e1003457.

Grabherr MG, Haas BJ, Yassour M, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, e1000695.

Hedrick P (2011) *Genetics of populations*. Jones and Bartlett Publishers, Sudbury, Massachusetts, USA.

Holdren C, Ehrlich P (1981) Long range dispersal in checkerspot butterflies: transplant experiments with Euphydryas gillettii. *Oecologia*, **50**, 125–129.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Lassmann T, Hayashizaki Y, Daub CO (2009) TagDust-a program to eliminate artifacts from next generation sequencing data. *Bioinformatics*, **25**, 2839–2840.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Li H, Handsaker B, Wysoker A, *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Locke DP, Hillier LW, Warren WC, *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**, 529–533.

Lohmueller KE, Bustamante CD, Clark AG (2009) Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics*, **182**, 217–231.

Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic parameters. *Bioinformatics*, **25**, 2747–2749.

Lukić S, Hey J (2012) Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics*, **192**, 619–639.

Messer PW, Petrov DA (2013) Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 8615–8620.

Min XJ, Butler G, Storms R, Tsang A (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Research*, **33**, W677–W680.

Molina J, Sikora M, Garud N, *et al.* (2011) Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences*, **108**, 8351.

Oberhauser KS (1989) Effects of spermatophores on male and female monarch butterfly 641 reproductive success. *Behavioral Ecology and Sociobiology*, **25**, 237–246.

Papanicolaou A, Gebauer-Jung S, Blaxter ML, Owen McMillan W, Jiggins CD (2008) ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Research*, **36**, D582–D587.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Subramanian S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, **168**, 373–381.

Whitlock MC (2000) Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. *Evolution*, **54**, 1855–1861.

Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

Williams E (1988) Habitat and range of *Euphydryas gillettii* (Nymphalidae). *Journal of the Lepidopterists' Society*, **42**, 37–45.

Williams EH, Holdren CE, Ehrlich PR (1984) The life history and ecology of *Euphydryas gillettii* Barnes (Nymphalidae). *Journal of the Lepidopterists' Society*, **38**, 1–12.

Winnebeck EC, Millar CD, Warman GR (2010) Why does insect RNA look degraded? *Journal of insect science (Online)*, **10**, 159.

Wright S (1931) Evolution in mendelian populations. *Genetics*, **16**, 97–159.

R.C.M. helped design experiment, prepared libraries, performed all downstream analyses and prepared manuscript. N.R.G. helped design experiment, provided guidance on analyses and commented on the manuscript. J.L.K. provided guidance for analyses and commented on the manuscript. C.L.B. collected demographic data and provided expertise related to the study system as well as guidance on analyses and comments on the manuscript. D.A.P. developed research question, helped design experiment, provided guidance on analyses and comments on the manuscript.

Data accessibility

The *Euphydryas gillettii* transcriptome assembly and variant calls are uploaded as online supporting material. Annual mark–release–recapture records are included as a supplementary table. Raw sequence data have been deposited in the NCBI Sequence Read Archive (BioProject ID: PRJNA222514; BioSample: SAMN02381172-SAMN02381187; accession numbers: SRX367091, SRX367185-SRX367189, SRX367194, SRX367198-SRX367206). Scripts used to perform the transcriptome assembly, demographic analyses, and performance analyses are uploaded as online supporting material.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Results of fitting model A with synonymous SNP data.

Fig. S2 Allele frequency spectra normalized to number of total SNPs of each class (synonymous and nonsynonymous).

Table S1 Summary statistics from Trinity *de novo* transcriptome assemblies, with kmer lengths ranging from $k = 21$ to $k = 31$, odd.

Table S2 Reciprocal blast searches of the *Euphydryas gillettii* transcriptome assembly (56 536 unique contigs) to the protein databases of *Bombyx mori*, *Danaus plexippus*, and *Heliconius melpomene*.

Table S3 Reciprocal blast searches of the portion of the *Euphydryas gillettii* transcriptome assembly (2757 unique contigs) in which high-confidence SNPs were discovered to the protein databases of *Bombyx mori*, *Danaus plexippus* and *Heliconius melpomene*.

Table S4 Estimates of *E. gillettii* population size ($\hat{N}$) at the Gothic, CO main site.

Table S5 Results of regression model with population growth as the response variable and June maximum temperature as a predictor variable.