MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 2455–2473 doi: 10.1111/j.1365-294X.2010.04666.x

Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies

ZACHARIAH GOMPERT,* MATTHEW L. FORISTER,† JAMES A. FORDYCE,‡ CHRIS C. NICE,§ ROBERT J. WILLIAMSON¶ and C. ALEX BUERKLE*

**Department of Botany, Program in Ecology, University of Wyoming, Laramie, WY 82071, USA, †Department of Biology/MS 314, University of Nevada, Reno, NV 89557, USA, ‡Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN 37996, USA, §Department of Biology, Population and Conservation Biology Program, Texas State University, San Marcos, TX 78666, USA, ¶Department of Applied Biology and Biomedical Engineering and Department of Computer Science and Software Engineering, Rose-Hulman Institute of Technology, Terre Haute, IN 47803, USA*

Abstract

The distribution of genetic variation within and among populations is commonly used to infer their demographic and evolutionary histories. This endeavour has the potential to benefit substantially from high-throughput next-generation sequencing technologies through a rapid increase in the amount of data available and a corresponding increase in the precision of parameter estimation. Here we report the results of a phylogeographic study of the North American butterfly genus *Lycaeides* using 454 sequence data. This study serves the dual purpose of demonstrating novel molecular and analytical methods for population genetic analyses with 454 sequence data and expanding our knowledge of the phylogeographic history of *Lycaeides*. We obtained 341 045 sequence reads from 12 populations that we were able to assemble into 15 262 contigs (most of which were variable), representing one of the largest population genetic data sets for a non-model organism to date. We examined patterns of genetic variation using a hierarchical Bayesian analysis of molecular variance model, which provides precise estimates of genome-level ϕ_{ST} while appropriately modelling uncertainty in locus-specific ϕ_{ST}. We found that approximately 36% of sequence variation was partitioned among populations, suggesting historical or current isolation among the sampled populations. Estimates of pairwise genome-level ϕ_{ST} were largely consistent with a previous phylogeographic model for *Lycaeides*, suggesting fragmentation into two to three refugia during Pleistocene glacial cycles followed by post-Pleistocene range expansion and secondary contact leading to introgressive hybridization. This study demonstrates the potential of using genome-level data to better understand the phylogeographic history of populations.

Keywords: 454 pyrosequencing, analysis of molecular variance, hierarchical Bayesian models, hybridization, Pleistocene phylogeography

Received 16 October 2009; revision received 13 January 2010; accepted 3 February 2010

Introduction

The distribution of genetic variation within and among populations can provide fundamental insights into their demographic and evolutionary history (Slatkin 1987; Avise 2004; Nielsen 2005; Patterson *et al*. 2006). Genetic diversity within populations is affected by their effective population sizes, immigration from other populations and mutation rates (Wright 1931; Gillespie 2004; Hedrick 2005), whereas the extent that genetic diversity is partitioned among populations is affected by gene

Correspondence: Zachariah Gompert, Fax: +1 307 766 2851; E-mail: zgompert@uwyo.edu

flow among populations and the divergence time of populations from a common ancestral population (Wright 1931, 1943; Felsenstein 1982; Slatkin 1993; Hedrick 2005). Selection further influences the distribution of genetic variation within and among populations for specific genetic regions where the fitness effects of allelic variants differ (Wright 1931; Beaumont & Balding 2004; Hedrick 2005; Nielsen 2005; Barrett *et al.* 2008). Numerous population genetic models have been derived that use genetic marker data to estimate various subsets of parameters and summary statistics associated with these processes (e.g. Wright 1931; Excoffier *et al.* 1992; Pritchard *et al.* 2000; Nielsen & Wakeley 2001; Hein *et al.* 2005). Estimation of evolutionary and demographic parameters from DNA sequence data is particularly powerful, as models of mutation for sequence data have been thoroughly developed and the relationship among alleles can be inferred (Kimura 1980; Felsenstein 2004). However, technical and logistical constraints have precluded most population genetic studies involving DNA sequence variation from utilizing more than a few sequence markers (e.g. Forster *et al.* 1996; Taberlet *et al.* 1998; Hoarau *et al.* 2007; Gompert *et al.* 2008a), although several recent studies, particularly those involving model organisms, represent an exception to this generalization (e.g. Lee & Edwards 2008; Strasburg & Rieseberg 2008; Patin *et al.* 2009). Reliance on a small number of markers for population genetic inference is not ideal, as the information contained in each marker represents a single realization of an inherently stochastic evolutionary processes (Hein *et al.* 2005; Degnan & Rosenberg 2006; Forister *et al.* 2008). This stochasticity makes it difficult to obtain precise and reliable parameter estimates.

High-throughput next-generation DNA sequencing technologies alleviate logistical constraints on generating large multilocus sequence data sets for population genetic parameter estimation in non-model systems. These new technologies are capable of generating on the order of 10^9 bp of sequence data in a single run (Margulies *et al.* 2005; Bentley 2006; Mardis 2008; Ondov *et al.* 2008). Thus, these technologies will facilitate population genetic inference of evolutionary and demographic parameters from hundreds or even thousands of sequence regions from many individuals and populations. To date, most published studies in evolution, ecology and genetics utilizing next-generation sequencing have focused on transcriptome assembly and characterization (e.g. Vera *et al.* 2008; Hale *et al.* 2009; Kristiansson *et al.* 2009; Meyer *et al.* 2009), although next-generation sequence data have also been used to estimate substitution rates, nucleotide diversity and inter-specific nucleotide divergence (Novaes *et al.* 2008; Hahn *et al.* 2009, Kulathinal *et al.* 2009). This dramatic increase in data should result in refined, highly reliable parameter estimation. These new technologies also present their own suite of difficulties. For example, concern exists that these new sequencing technologies tend to have higher error rates than traditional Sanger sequencing, and could lead to overestimation of levels of molecular polymorphism, although a recent study by Harismendy *et al.* (2009) suggests that this may not be the case. Methods accounting for sequence error are already under development (Hellmann *et al.* 2008; Lynch 2008, 2009). Additionally, these technologies use 'shotgun' sequencing strategies that are likely to result in uneven sequence coverage among individuals and populations for specific genetic regions. In fact, many individuals and populations may be missing sequence data for a given genetic region. This issue is likely to be particularly important when Eukaryotic genomic DNA is used as a template because of the size and complexity of Eukaryote genomes (van Orsouw *et al.* 2007). Statistical methods addressing these issues are being developed rapidly (e.g. Hellmann *et al.* 2008; Lynch 2008; Novaes *et al.* 2008; Lynch 2009) and molecular methods for targeted resequencing on next-generation platforms may further ameliorate these problems (e.g. Okou *et al.* 2007; Gnirke *et al.* 2009; Hodges *et al.* 2009). New methods for population genetic inference will need to utilize as much information as possible from the large quantity of data generated by next-generation sequencing platforms, while appropriately modelling uncertainty due to uneven coverage and missing data.

In the current study we use 454 sequence data from a large number of sequenced fragments in conjunction with a novel modelling approach to investigate the population genetic structure of North American *Lycaeides* butterflies (Lepidoptera: Lycaenidae). This study is meant to serve both as a test case for these new methods and to increase our understanding of the evolutionary history of *Lycaeides* butterflies. To generate sequence data useful for population genetic inference, we create a reduced complexity genomic DNA library for *Lycaeides* using an AFLP-based technique, and attach 10-bp sequence tags (Multiplex Identifier or MID barcodes; 454 Life Sciences Corp. 2009) to template fragments allowing individual sequences to be assigned to populations. We investigate the genetic structure of *Lycaeides* using a hierarchical Bayesian model for analysis of molecular variance (AMOVA). Unlike coalescent methods, AMOVA does not estimate evolutionary process parameters. Nonetheless, ϕ-statistics from AMOVA are commonly used, as they do not require assumptions regarding the evolutionary history of populations and can provide substantial phylogeographic insights (Holsinger & Weir 2009). Hierarchical Bayesian models are well suited for the analysis of next-generation sequence

data as they provide a robust framework for inferring genome-level parameters from locus-specific data while appropriately modelling uncertainty, including the uncertainty arising from missing data and uneven coverage (Gelman *et al.* 2004; Guo *et al.* 2009). The molecular and analytical techniques we use have limitations (see Discussion) and will certainly be improved upon in future. Nonetheless, we believe that this study details a novel approach to obtaining the best possible estimates of key population genetic parameters from next-generation sequence data. Finally, our previous knowledge of *Lycaeides* butterflies makes this study system an ideal test case for these novel methods.

The phylogeographic history of North American *Lycaeides* butterflies is complex and includes periods of geographic isolation and subsequent hybridization (Nice & Shapiro 1999; Nice *et al.* 2005; Gompert *et al.* 2006a,b, 2008a,b; Lucas *et al.* 2008). This complexity has led to uncertainty in taxonomic designations within this group, with two to five nominal species recognized or proposed and many more recognized subspecies (Nabokov 1949; Scott 1986; Guppy & Shepard 2001; Gompert *et al.* 2006a,b). Although we utilize species designations in this manuscript (Table 1), we do so primarily to facilitate communication, and not because we believe that these named lineages necessarily represent reproductively isolated entities. Instead, we believe that these lineages are best thought of as a species complex, and treat them as such in our analyses. *Lycaeides* is a circumpolar holarctic genus that probably colonized North America within the last two to four million years (Gompert *et al.* 2008a). Previously published molecular data suggest that diversification of *Lycaeides* within North America occurred within the last few hundred thousand years, coinciding with fragmentation during Pleistocene glacial cycles (Nice *et al.* 2005; Gompert *et al.* 2008a). These data are consistent with fragmentation of populations into at least three glacial refugia: one in the western USA serving as the source of *L. idas* populations, one in the central USA serving as the source *L. melissa* populations and one in the eastern USA serving as the source of Karner blue (*L. melissa*

Table 1 Sample data for specimens used for 454 pyrosequencing

MID no.	Taxon	Locality	Latitude	Longitude	Sample size	Classification
1	*L. idas**	Yuba Gap, CA	39°29′34″N	120°35′39″W	3f 2m	W
1	*L. idas**	Leek Springs, CA	38°37′59″N	120°14′24″W	3f 2m	W
1	*L. idas**	Trap Creek, CA	39°22′43″N	120°40′27″W	3f 2m	W
3	*L. idas**	Cave Lake, CA	41°58′46″N	120°12′25″W	6f 9m	W
4	*L. idas**	Marble Mts., CA	41°49′40″N	122°44′52″W	2f 5m	W
4	*L. idas**	Mt. Ashland, OR	42°04′52″N	122°43′16″W	0f 8m	W
5	*L. sp* 'alpine hybrid'	Carson Pass, CA	38°42′ 47″N	120°01′17″W	4f 4m	W × C
5	*L. sp* 'alpine hybrid'	Mt. Rose, NV	39°19′ 21″N	119°55′48″W	4f 3m	W × C
6	*L. sp* 'alpine hybrid'	County Line Hill, CA	37° 27′51″N	118°11′35″W	1f 14m	W × C
7	*L. sp* 'Warner entity'	Eagle Peak, CA	41°15′ 38″N	120°12′11″W	4f 11m	W × C
8	*L. melissa*	Gardnerville, CA	38°48′54″N	119°46′44″W	3f 2m	C
8	*L. melissa*	Verdi, NV	39°03′01″N	119°55′48″W	3f 2m	C
8	*L. melissa*	Sierravalley, CA	39°37′48″N	120°21′40″W	3f 2m	C
9	*L. melissa*	Beckwourth Pass, CA	39°47′35″N	120°06′38″W	4f 4m	C
9	*L. melissa*	Montague, CA	41°46′21″N	122°28′38″W	2f 5m	C
10	*L. melissa*	Brandon, SD	43°36′29″N	96°34′39″W	3f 2m	C
10	*L. melissa*	Victor, ID	43°39′32″N	111°06′41″W	3f 2m	C
10	*L. melissa*	Indian Bathtubs, WY	41°12′14″N	106°46′17″W	2f 3m	C
11	*L. melissa* 'Karner'	Fort McCoy, WI	43°47′ 59″N	90°49′59″W	0f 5m	E
11	*L. melissa* 'Karner'	Necedah, WI	44°04′00″ N	90°11′20″W	0f 5m	E
11	*L. melissa* 'Karner'	Saratoga, NY	43°03′24″ N	73°48′47″W	0f 5m	E
12	*L. idas*	Blacktail Butte, WY	43°38′17″N	110°40′55″W	3f 2m	W × C
12	*L. idas*	Riddle Lake, WY	44°21′42″N	110°32′48″W	2f 3m	W × C
12	*L. idas*	Jardine, MT	45°04′29″N	110°38′01″W	3f 2m	W × C
13	*L. idas*	Prospect Creek, AB	52°58′01″N	117°23′02″W	1f 7m	W
13	*L. idas*	Brule, AB	53°16′59″N	117°52′06″W	1f 7m	W

Populations are referred to by their MID numbers. Classifications correspond to hypothesized glacial refugia: western (W), central (C), eastern (E) and populations hypothesized to contain admixed individuals derived from western and central refugial populations (W × C). See text for more details.

*Designated *L. anna* by Guppy & Shepard (2001).

samuelis) populations (Nice *et al.* 2005). A fourth refugium might have existed in Alaska. Previously published data suggest that post-Pleistocene range expansion led to secondary contact among formerly isolated refugial lineages in the Sierra Nevada, Rocky Mountains and the eastern USA (Nice *et al.* 2005; Gompert *et al.* 2006a,b, 2008b; Lucas *et al.* 2008). These data indicate that hybridization has occurred in these regions of secondary contact with various outcomes. For example, in the eastern USA mitochondrial introgression has occurred with little or no nuclear introgression, perhaps driven by an association between mitochondrial haplotype and *Wolbachia* infection status (Gompert *et al.* 2006b, 2008b; Lucas *et al.* 2008; Nice *et al.* 2009). Conversely, secondary contact in the Rocky Mountains and Sierra Nevada (as well as other nearby ranges, e.g. the Warner and White mountains) has involved substantial admixture and the establishment of a homoploid hybrid species in the high-altitude alpine habitat of the Sierra Nevada (Gompert *et al.* 2006a, 2008b; Lucas *et al.* 2008).

The historical phylogeographic model described above, which is based primarily on geographic patterns of morphological and genetic variation (mtDNA, AFLPs and sequence data from a few nuclear genes), provides expectations that can be investigated with 454 pyrosequence data. Specifically, we make the following predictions regarding the population genetic structure of North American *Lycaeides*: (i) a significant proportion of 454 sequence variation will be partitioned among populations, (ii) molecular differentiation will be lower for populations arising from the same ancestral refugium, than from populations arising from different glacial refugia and (iii) Molecular differentiation between putatively admixed populations in regions of secondary contact and nearby parental populations should be lower than molecular differentiation between their parental populations. Testing these predictions using 454 sequence data will refine our understanding of the phylogeographic history of North American *Lycaeides* butterflies, and provide an illustrative example of molecular and analytical methods that can be used to facilitate population genetic inference from next-generation sequence data.

Methods

We sampled 15 adult *Lycaeides* butterflies from each of 12 populations composed of one to three sampling localities (Fig. 1, Table 1). These populations represent most major North American lineages and correspond to genetically cohesive entities that were defined based on previous morphological and phylogeographic studies of *Lycaeides* (e.g. Nice *et al.* 2005; Gompert *et al.* 2006a, 2008a,b; Lucas *et al.* 2008). We isolated and purified DNA from each of the 180 sampled butterflies from approximately 10 mg of thoracic tissue using: (i) standard methods from Brookes *et al.* (1997), or (ii) Qiagen's DNeasy 96 Blood and Tissue Kit (Qiagen Inc.) in accordance with the manufacturer's recommended protocol. The concentration and purity of the DNA samples were determined using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific). Subsequently, all DNA samples were diluted to achieve DNA concentrations of 10 ng/μL. The 15 individual DNA samples for each population were combined in equal amounts to produce 12 pooled population samples. These 12 samples served as templates for subsequent reactions.

We reduced the complexity of the genomic DNA templates using an AFLP-based procedure similar to the CRoPS method described by van Orsouw *et al.* (2007). First, we digested 60 ng of genomic DNA for each of the 12 population samples with two restriction endonucleases (*Eco*RI and *Mse*I) and ligated the digested fragments to double stranded adaptor oligos in a single step at 37 °C for 18 h. We then diluted the ligated DNA fragments 10-fold with 0.1x TE buffer. An initial PCR amplification of this product with preselective AFLP primers (EcoRI and MseI-C) was performed in BioRad's MyCycler (BioRad Laboratories) with 30 s at 98 °C, 30 cycles of 20 s at 98 °C, 30 s at 56 °C and 120 s at 72 °C followed by a final extension for 10 min at 72 °C. Amplification was achieved using iProof high-fidelity DNA polymerase (BioRad Laboratories) to reduce PCR error that could negatively affect 454 pyrosequencing results. Amplicons from the PCR reaction were separated on a 2% agarose gel and those fragments between approximately 400 and 550 bp were excised from the gel and purified using the GENECLEAN Turbo DNA purification kit (MP Biomedicals, LLC). We believed that fragments in this range would be ideally suited for 454 pyrosequencing given the distribution of read lengths obtainable on the 454 GS XLR70 Titanium platform. To obtain a sufficient quantity of template and to ensure a sufficient number of unique reads for population genetic analyses, we performed a second PCR amplification using the original digested DNA template under conditions identical to those described above. PCR fragments produced from this second amplification were also purified from 2% agarose gel; however, for this second amplification we excised and purified amplicons between approximately 200 and 550 bp. The increased range of fragments for this second PCR amplification was chosen to increase the total number of unique amplicons. The two PCR products of purified amplicons for each population DNA sample were combined into a single template for subsequent procedures.

A final PCR amplification was performed for each of the 12 amplicon samples to increase the amplicon con-

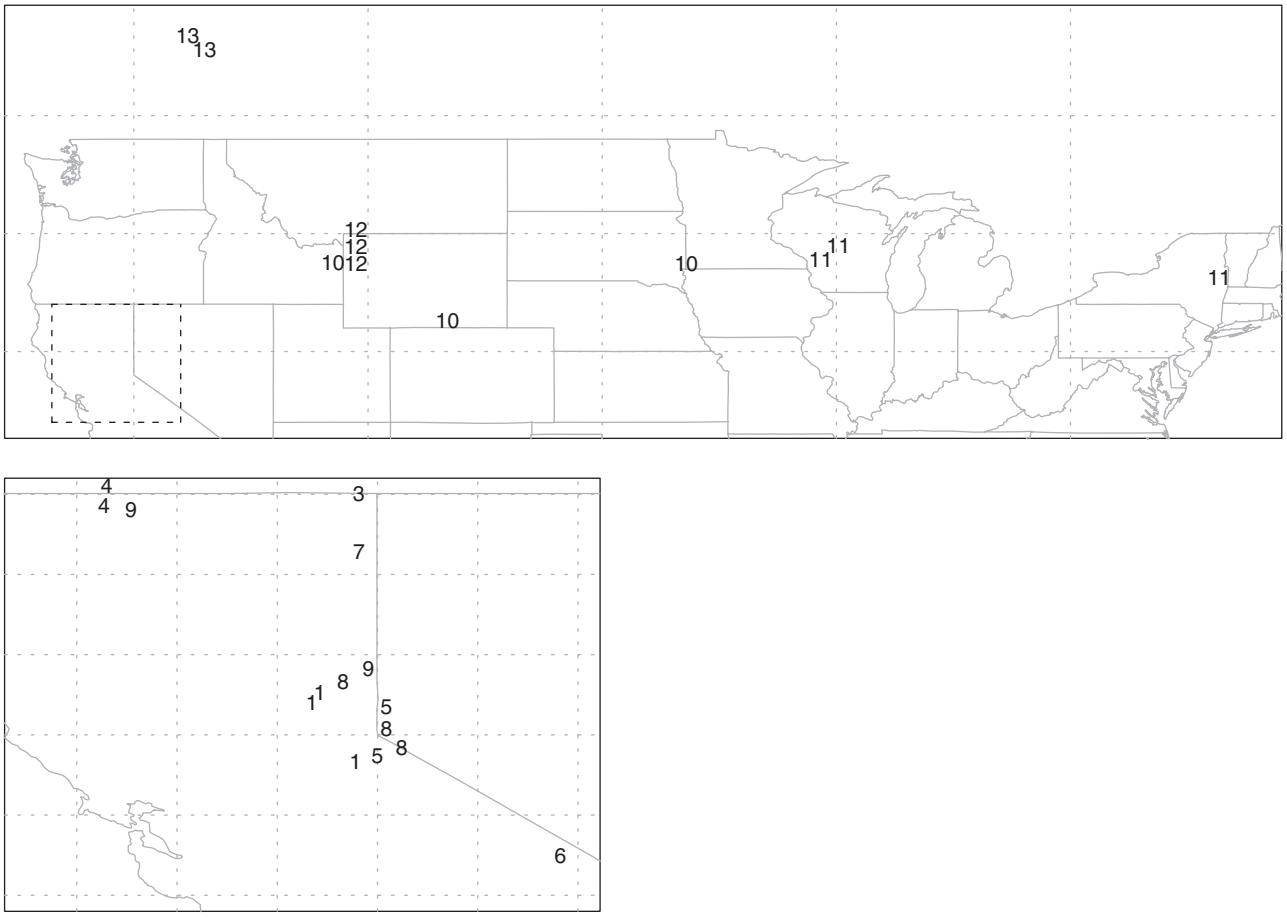

Fig. 1 Map of localities for sampled *Lycaeides*. Numbers correspond to MID barcode labels used for each population (some populations are composed of multiple localities). The dashed line designates the geographic region shown in more detail below the main plot.

Table 2 MID barcodes and PCR primers used for population identification

MID no.	MID EcoRI primer	MID MseI primer
1	5′-ACGAGTCGCT GACTGCGTACCAATTC-3′	5′-ACGAGTCGCT GATGAGTCCTGAGTAAC-3′
3	5′-AGACGCACTC GACTGCGTACCAATTC-3′	5′-AGACGCACTC GATGAGTCCTGAGTAAC-3′
4	5′-AGCACTGTAG GACTGCGTACCAATTC-3′	5′-AGCACTGTAG GATGAGTCCTGAGTAAC-3′
5	5′-ATCAGACACG GACTGCGTACCAATTC-3′	5′-ATCAGACACG GATGAGTCCTGAGTAAC-3′
6	5′-ATATCGCGAG GACTGCGTACCAATTC-3′	5′-ATATCGCGAG GATGAGTCCTGAGTAAC-3′
7	5′-CGTGTCTCTA GACTGCGTACCAATTC-3′	5′-CGTGTCTCTA GATGAGTCCTGAGTAAC-3′
8	5′-CTCGCGTGTC GACTGCGTACCAATTC-3′	5′-CTCGCGTGTC GATGAGTCCTGAGTAAC-3′
9	5′-TAGTATCAGC GACTGCGTACCAATTC-3′	5′-TACTATCAGC GATGAGTCCTGAGTAAC-3′
10	5′-TCTCTATGCG GACTGCGTACCAATTC-3′	5′-TCTCTATGCG GATGAGTCCTGAGTAAC-3′
11	5′-TGATACGTCT GACTGCGTACCAATTC-3′	5′-TGATACGTCT GATGAGTCCTGAGTAAC-3′
12	5′-TACTGAGCTA GACTGCGTACCAATTC-3′	5′-TACTGAGCTA GATGAGTCCTGAGTAAC-3′
13	5′-CATAGTAGTG GACTGCGTACCAATTC-3′	5′-CATAGTAGTG GATGAGTCCTGAGTAAC-3′

centration and label the amplicons with sample-specific MID barcodes. For each sample a distinct pair of primers was used with a core sequence identical to that used for PCR described above, but a unique 10-bp sequence added to the 5′-end for subsequent population identification (Table 2). Amplification was performed as described above and again utilized iProof high-fidelity DNA polymerase (BioRad Laboratories) to minimize

PCR error. The resulting PCR product was purified using the GENECLEAN Turbo DNA purification kit (Biomedicals, LLC) following the manufacturer's protocol. The concentration of PCR product for each sample was determined using a NanoDrop D-1000 spectrophotometer (Thermo Scientific) and each sample was subsequently diluted to a concentration of 100 ng/μL. All 12 population samples were then combined in equal amounts to yield a single 100 μL product sample with an amplicon concentration of 100 ng/μL, which served as the template for 454 pyrosequencing.

454 pyrosequencing was performed by SeqWright DNA Technology Services (Houston, TX, USA) using the 454 GS XLR70 Titanium platform. Pyrosequencing was performed as described by Margulies *et al.* (2005) following standard protocols provided by 454 Life Sciences. Specifically, to prepare a library for 454 pyrosequencing, fragments from the labelled amplicon sample were blunt ended by enzymatic polishing using T4 DNA polymerase. Adaptors were then ligated to the polished amplicon fragments and single-stranded fragments were individually immobilized on capture beads. Emulsion PCR was then accomplished by clonally amplifying individual immobilized fragments within emulsion droplets. Beads containing amplified single-stranded DNA fragments were packed on 1/2 of a 70×75-mm^2 PicoTiterPlate for pyrosequencing.

Sequence assembly and characterization

We assembled the sequences generated from 454 pyrosequencing into contigs using the SeqMan NGen sequence assembler v2.0.0 (DNASTAR) software. Prior to sequence assembly MID barcode sequences were removed from the reads and appended to the read names for later identification. Assembly was conducted with a mer match size of 75 bp, mer spacing of 20 bp and a minimum mer match percentage of 92%. The repeat handling option was used while assuming a maximum fixed coverage of 50× and a match repeat per cent of 150. We set the nucleotide mismatch penalty to 15 and the gap penalty to 150. Quality scores from the 454 run were provided to the SeqMan NGen assembly software and used to trim low-quality regions from the reads. Specifically, reads with an average quality score of 14 or less over a 30 bp window were flagged as low-quality reads and removed. We also provided the core AFLP primer sequences to the assembly software so that these regions could be trimmed from the reads during the assembly process. The parameters used for sequence assembly were selected following initial experimentation with a range of parameters and were chosen to provide the greatest number of quality contigs. The complete set of parameters used for SeqMan NGen assembly is available from the authors upon request.

To characterize the sequences we obtained from 454 pyrosequencing, we conducted BLAST searches of our assembled contigs and unassembled reads against the UniRef50 database (Suzek *et al.* 2007). The aim of the BLAST search was primarily to determine the proportion of our contigs and sequences that were associated with genes, not to provide a complete annotation of these sequences. The UniRef50 database consists of clustered sets of protein coding sequences from UniProt Knowledgebase (Bairoch *et al.* 2009) that share ≥50% sequence identity. BLAST searches were performed with the BLASTN algorithm at an *e*-value threshold of 10^{-11}. Results from BLAST searches were parsed and summarized with a series of custom Perl scripts (available from ZG upon request). Gene ontology (GO) classifications for unique UniRef50 accessions with significant similarity to our assembled contigs and unassembled reads were obtained using blast2GO (Conesa *et al.* 2005). GO terms were mapped to the UniRef50 accessions by accessing annotation files from the GO consortium on the blast2GO server. The GO terms were then assigned to our accessions using the blast2GO annotation rule with default parameters, except that the *e*-value threshold was set to 10^{-11}. Tests of equal proportions and Fisher's exact tests were conducted to compare BLAST results and GO classifications between the contigs and unassembled reads. All general statistical analyses were conducted using the R software environment for statistical computing (R Development Core Team 2009).

Population genetic analyses

Prior to assessing patterns of population genetic structure, we characterized the genetic diversity present within the 454 data. First, we calculated the proportion of variable sites for each of the assembled contigs (15 262 contigs). We treat this as an empirical measure of the proportion of sites that varied in our data, not as an estimate of the actual proportion of variable sites in the sampled *Lycaeides* populations. The proportion of variable sites was determined with and without the inclusion of insertion–deletion polymorphisms. Next we estimated SNP allele frequencies within each population for all nucleotide positions where more than one base was observed. Maximum likelihood estimates of SNP allele frequencies were obtained using a model similar to a model proposed by Lynch (2009) but adapted for population-level data. Our method of SNP allele frequency estimation assumed that the frequency of amplicons for each population sample was not affected by PCR (i.e. we did not account for differential amplification of alternative SNP alleles). For each SNP site we

assumed that no more than two nucleotides were segregating within a population. We believe that this assumption is valid as 97.6% of the SNPs we detected had two alleles, as is the case for many empirical data sets (Lynch 2007). For each SNP we defined counts for the major allele (the most common nucleotide observed, n_1), the minor allele (the second most common nucleotide observed, n_2) and erroneous reads (the sum of the least common nucleotides, n_ε). We then assumed that the counts followed a multinomial distribution incorporating three parameters: p_1, p_2 ($p_2 = 1 - p_1$) and ε reflecting the frequency of the major allele in the sample, the frequency of the minor allele in the sample and the probability of a sequencing error respectively. We assumed that read errors were equally likely to occur for both SNP alleles and that an error was equally likely to produce a read of any of the other three nucleotides. We allowed the probability of sequencing errors (ε) to vary among SNP loci. Furthermore, we assumed that p_1, the major allele frequency for the sampled individuals, followed a beta-distribution with parameters $\alpha = v\pi_1 + 1$ and $\beta = v(1 - \pi_1) + 1$, where v was twice the number of individuals sampled from each population (i.e. the number of gene copies sampled) and π_1 was the population allele frequency of the n_1 allele. The population frequency of n_2 was simply $1 - \pi_1$. This model yields the following likelihood function:

$$P(n_1,n_2,n_\varepsilon|n,p_1,\varepsilon,\pi_1,v) = \left[p_1(1-\varepsilon) + (1-p_1)\left(\frac{\varepsilon}{3}\right)\right]^{n_1} \left[(1-p_1)(1-\varepsilon) + p_1\left(\frac{\varepsilon}{3}\right)\right]^{n_2} \left[\frac{2\varepsilon}{3}\right]^{n_\varepsilon} \left[\frac{\Gamma(v+2)}{\Gamma(v\pi_1+1)\Gamma(v(1-\pi_1)+1)} p_1^{v\pi_1}(1-p_1)^{v(1-\pi_1)}\right] \quad (eqn1)$$

where n is the total number of sequences from a population at a specific site. Note, the first bracketed term gives the probability of observing the n_1 allele (which is the sum of the probability of having the n_1 allele and not having a sequencing error and the probability of having the n_2 allele that is converted to the n_1 allele because of a sequencing error) and the second bracketed term gives the probability of observing the n_2 allele. The third bracketed term gives the probability of a sequencing error giving rise to an allele not segregating in the population. Finally, the last bracketed term is a probability density function for a beta-distribution and provides the probability of the sample major allele frequency (p_1) if v gene copies were sampled from a population with a major allele frequency π_i. The mode of a beta-distribution is $(\alpha - 1)/(\alpha + \beta - 2)$. Thus, the maximum likelihood estimate of π_1 ($\hat{\pi_1}$) is equal to the maximum likelihood estimate of p_1 ($\hat{p_1}$),

$$\hat{p_1} = \frac{v\hat{\pi_1}}{v\hat{\pi_1} + v(1-\hat{\pi_1})} = \hat{\pi_1}.$$

Maximum likelihood estimates of p_1 and ϵ were obtained by a grid search over possible values based on the first three terms in Eqn 1; the maximum likelihood estimate of p_1 was then equated with the maximum likelihood estimate of π_1. A more complicated numerical evaluation of Eqn 1 would be necessary to obtain confidence intervals for p_1, ϵ and π_1. We treated gaps as missing data and only estimated SNP allele frequencies for sites with five or more sequences from a given population. Moreover, we excluded contigs with annotations suggesting bacterial sources and contigs more than 700 bp in length (the latter generally corresponded to repetitive regions of DNA potentially containing sequences from non-orthologous genes). These criteria excluded 688 of the 15 262 contigs. These analyses were conducted using custom Perl and R scripts (available from ZG upon request). Simulations suggested that this method for SNP allele frequency estimation yields unbiased estimates under many circumstances with increased precision as the number of sequence reads increases (see Document S1). However, this is not true when the true major allele frequency is close to 0.5 as estimates of the major allele frequency are constrained to be ≥0.5.

We next estimated ϕ_{ST} (the proportion of molecular variation partitioned among populations) in an AMOVA framework to determine whether there was evidence of population genetic structure in our 454 data. We used a hierarchical Bayesian model to estimate ϕ_{ST} for each locus, assuming that the locus-specific ϕ_{ST} values were drawn from a common distribution with a mean equal to the genome-level ϕ_{ST}. These estimates were made while accounting for uncertainty in population haplotype frequencies. We assumed that the vector haplotype count data (x_{ij}) for each locus (i) and population (j) were distributed as a multivariate Pólya distribution (also known as a Dirichlet compound multinomial distribution) with a parameter vector α_{ij} giving the population haplotype frequencies. The multivariate Pólya distribution describes a set of counts (e.g. haplotype counts) drawn from a multinomial distribution with a parameter vector p (i.e. sample haplotype frequencies) drawn from a Dirichlet distribution with parameter vector α (e.g. population haplotype frequencies). The probability density function for the Pólya distribution is derived by integrating over all possible values for p (i.e. all possible sample haplotype frequencies). Thus, using this probability distribution allowed us to estimate population haplotype frequencies while accounting for error associated with sampling individuals from populations and sequences from the sampled

individuals. This model specification yields the following likelihood function:

$$P(\mathbf{X}|\mathbf{A},\mathbf{n},v) = \prod_i \prod_j \frac{n_{ij}!}{\prod_k (x_{ijk}!)} \frac{\Gamma(\sum_k v\alpha_{ijk}+1)}{\Gamma(n_{ij}+\sum_k v\alpha_{ijk}+1)} \prod_k \frac{\Gamma(x_{ijk}+v\alpha_{ijk}+1)}{\Gamma(v\alpha_{ijk}+1)} \quad (eqn2)$$

where Γ is the gamma function, n_{ij} is the number of 454 sequences for locus i from population j, v is the number of gene copies sampled from each population (i.e. 2× the number of sampled individuals) and $\mathbf{X}$ and $\mathbf{A}$ are three-dimensional matrices (locus $i \times$ population $j \times$ haplotype k) containing the observed haplotype counts (x_{ijk}) and population haplotype frequencies (α_{ijk}) respectively. We assigned a conditional prior to $\mathbf{A}$:

$$P(\mathbf{A}|\mu_\phi,\sigma_\phi,\mathbf{D}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_\phi^2}} \exp\left[\frac{-(\phi_{\mathrm{ST}_i}-\mu_\phi)^2}{2\sigma_\phi^2}\right] \quad (eqn3)$$

where μ_ϕ corresponds to ϕ_{ST} at the genome level, σ_ϕ (i.e. $\sqrt{\sigma_\phi^2}$) is the average deviation of each locus-specific ϕ_{ST} from the genome level ϕ_{ST}. The variable ϕ_{ST_i} in Eqn 3 denotes ϕ_{ST} for locus i calculated from α_i and d_i following Excoffier *et al.* (1992). This specification of the conditional prior does not correspond to a standard probability distribution for $\mathbf{A}$ but is equivalent to assuming that the locus-specific ϕ_{ST} are distributed $N(\mu_\phi,\sigma_\phi^2)$ with $\mathbf{D}$ fixed and known without error (given a fixed distance matrix, any set of allele frequencies results in a specific value for ϕ_{ST}). Finally, we assigned uninformative hyper-priors to μ_ϕ and σ_ϕ, such that $\mu_\phi \sim N(\mu=0,\sigma^2=10^6)$ and $\sigma_\phi \sim \mathrm{beta}(\alpha=1,\beta=1)$. Combining these functions resulted in the following hierarchical Bayesian model:

$$P(\mathbf{A},\mu_\phi,\sigma_\phi|\mathbf{X},\mathbf{D},\mathbf{n},v) \propto P(\mathbf{X}|\mathbf{A},\mathbf{n},v) P(\mathbf{A}|\mu_\phi,\sigma_\phi,\mathbf{D}) P(\mu_\phi) P(\sigma_\phi). \quad (eqn4)$$

Our treatment of $\mathbf{D}$ in Eqn 4 is similar to the general treatment of covariates in Bayesian regression models (Gelman *et al.* 2004).

Unlike our model for estimating SNP allele frequencies, this Bayesian model does not account for sequencing error. It was not possible to account for sequence error using haplotype data in the same manner as we did for SNP data (i.e. it would not be appropriate to assume that only two haplotypes are segregating within or among populations). Although this is a potential limitation, we believe that the effect of not modelling sequence error should be minor, as our estimates of sequence error (excluding insertion–deletion polymorphisms; see the Results section) suggest that such errors were rare. Finally, this method assumes that all sequences in a contig represent orthologous genes. The parameters we chose for sequence assembly and the exclusion of long contigs from analyses should minimize contigs containing non-orthologous sequences. Nonetheless, some contigs may include non-orthologous sequences and the inclusion of non-orthologous sequences in a contig could negatively affect our estimates of ϕ_{ST}.

We estimated ϕ_{ST} using 1570 contigs with greater than 10× coverage (over all populations) and that were less than 700 bp in length. We did not include contigs with annotations suggesting bacterial origin. For each contig we removed all sites polymorphic for insertions–deletions and trimmed contigs to obtain a rectangular matrix without missing data. From these trimmed contigs we counted the number of occurrences of each haplotype at each locus for each population ($\mathbf{X}$) and computed pairwise molecular distance matrices for all haplotypes in each contig ($\mathbf{D}$). Distances were calculated as the number of sites that differed between a pair of haplotypes using functions provided in the R package *ape* (Paradis *et al.* 2004). We estimated locus-specific and genome-level ϕ_{ST} for all 12 populations following the hierarchical Bayesian model described above and using Markov chain Monte Carlo (MCMC). We employed a Metropolis–Hastings MCMC algorithm, with the proposal distribution for μ_ϕ centred on the estimate of μ_ϕ from the previous iteration (i.e. random walk algorithm) and fixed densities for the proposal distributions of σ_ϕ and each haplotype frequency vector α_{ij} (i.e. independence chain algorithm). We used a burn-in of 1000 MCMC iterations followed by 24 000 additional iterations for parameter estimation. Sample paths for α_{ij}, ϕ_{ST}, μ_ϕ and σ_ϕ were monitored to ensure adequate chain mixing and convergence of the MCMC chain on the stationary distribution ($P(\mathbf{A},\mu_\phi,\sigma_\phi|\mathbf{X},\mathbf{D})$). The latter was also assessed by running multiple MCMC chains from different starting points. MCMC required approximately 6 hours on a Macintosh computer with 2×2.8 GHz Quad-Core Intel Xeon processors and 4 GB 800 MHz DDR2 FB-DIMM memory. This MCMC algorithm was programmed in C using the GNU Scientific Library (Galassi *et al.* 2009). Source code for this model is available from ZG upon request and a user-friendly compiled binary for this analysis is forthcoming.

In addition to estimating ϕ_{ST} for North American *Lycaeides* as a whole, we estimated pairwise ϕ_{ST} for all pairs of the 12 sampled populations. Pairwise ϕ_{ST} estimates give the proportion of molecular variation partitioned between a pair of populations and can serve as a measure of molecular differentiation between pairs of populations over short evolutionary timescales

(Reynolds *et al.* 1983; Slatkin 1995). Pairwise ϕ_{ST} were estimated using the hierarchical Bayesian model described above but with haplotype data from pairs of populations. Our estimate of **A** allowed for non-zero haplotype frequencies for all haplotypes observed, not just those haplotypes observed in a given pair of populations. For all pairwise comparisons, we ran an MCMC chain for a 1000 iteration burn-in followed by 24 000 iterations for parameter estimation. Similar to the analysis of the full data set, we monitored sample paths for α_{ij}, ϕ_{ST}, μ_ϕ and σ_ϕ to ensure convergence on the stationary distribution. We used non-metric multidimensional (NMDS) scaling to visualize population genetic structure based on estimates of pairwise ϕ_{ST}. NMDS is a numerical ordination technique suitable for any pairwise dissimilarity measure and is appropriate for depicting geographic patterns of genetic structure (Lessa 1990; Venables & Ripley 2002). This method was chosen instead of tree-based clustering methods to avoid implying that the sampled *Lycaeides* populations are related in a bifurcating manner. We used NMDS to identify three dimensions that best retained the pairwise distances between populations based on pairwise ϕ_{ST} estimates (more specifically, the median from the posterior distribution for each pairwise ϕ_{ST}). This ordination was performed in R using the *MASS* package (Venables & Ripley 2002).

Results

Sequence assembly and characterization

We obtained 341 045 reads from sequencing our barcode-labelled template with a 1/2 PicoTiterPlate run on the 454 GS XLR70 Titanium platform. Files containing these sequence reads and quality scores have been submitted to the NCBI Short Read Archive (accession SRA010351). The mean read length (excluding the MID barcodes) was 242.18 bp (SD 134.97 bp; Fig. 2a). The mean number of reads per MID barcode (i.e. population DNA sample) was 28 420, with the fewest reads (19 110) obtained from MID8 and the most reads (39 630) obtained from MID12 (Fig. 2b). Of the 341 045 reads we obtained, we were able to assemble 134 772 reads into 15 262 contigs, leaving 206 273 unassembled sequences. These unassembled sequences represented a mix of reads not matching any other reads we obtained and low-quality reads excluded from the assembly (i.e. reads containing highly repetitive elements preventing their clustering within any one contig). The average quality score for the assembled sequences was 34, compared with 17 for the unassembled sequences. The mean contig length from the assembly was 310.39 bp (SD 152.89 bp; Fig. 2c). The number of reads per contig varied substantially among contigs but had a mean of 8.83 (SD 55.85), yielding an average coverage depth of 8.44 reads per site. The similarity between the number of reads per contig and average coverage depth indicates that most reads extended over an entire contig. The distribution of reads per contig was highly skewed with 10 545 contigs containing fewer than five reads (Fig. 2d). Nonetheless, contigs with many reads were also assembled. For example, there were 1995 contigs containing more than 10 reads and 385 contigs containing more than 50 reads. The mean number of assembled reads per MID barcode was 11 231 reads, with the fewest assembled reads (4151) for MID12 and the most assembled reads (18 185) for MID4 (Fig. 2e). MID12 had relatively few reads assembled under many different combinations of assembly parameters (results not shown). Excluding MID12 there was a significant correlation between the number of reads per MID barcode and the number of assembled reads per MID barcode (r=0.8280, P=0.0016; Fig. 2f); however, this correlation was not significant when MID 12 was included (r=0.2998, P=0.3438).

Of the 15, 262 contigs assembled, 2238 (14.7%) had significant hits to clusters in the UniRef50 database at an *e*-value threshold of 10^{-11}. The same was true for 21 974 of our 206 273 (10.7%) unassembled reads. Multiple contigs or reads had best BLAST hits to the same UniRef50 accessions. This may include instances where different genes in a gene family were sufficiently similar to be collapsed into a single accession in the UniRef50 database. Accounting for this redundancy, our assembled contigs and unassembled reads had significant BLAST hits to 828 and 1512 unique gene accessions in the UniRef50 database respectively. Both the proportion of sequences with significant matches to UniRef50 accessions and the proportion of sequences matching unique UniRef50 accessions was significantly greater for the assembled contigs than the unassembled reads (proportion tests: χ^2=234.44, $P<2.2\times10^{-16}$ and χ^2=2989.12, $P<2.2\times10^{-16}$ respectively). Most of the UniRef50 accessions corresponded to sequences obtained from arthropods. For example, the four most common taxonomic identifications associated with these gene accessions were *Tribolium castaneum* (Coleoptera, 69 hits), *Acyrthosiphon pisum* (Hemiptera, 41 hits), Coelomata (39 hits) and *Nasonia vitripennis* (Hymenoptera, 38 hits) for the assembled contigs and *T. castaneum* (98 hits), *A. pisum* (86 hits), *N. vitripennis* (70 hits) and Endopterygota (all holometabolic insects, 58 hits) for the unassembled reads. However, the BLAST results suggest that a portion of our 454 sequence data represents contaminant bacterial DNA, most likely α-proteobacteria in the genus *Wolbachia*. Specifically, 24 and 33 of the unique gene accessions hit by our BLAST searches were sequences

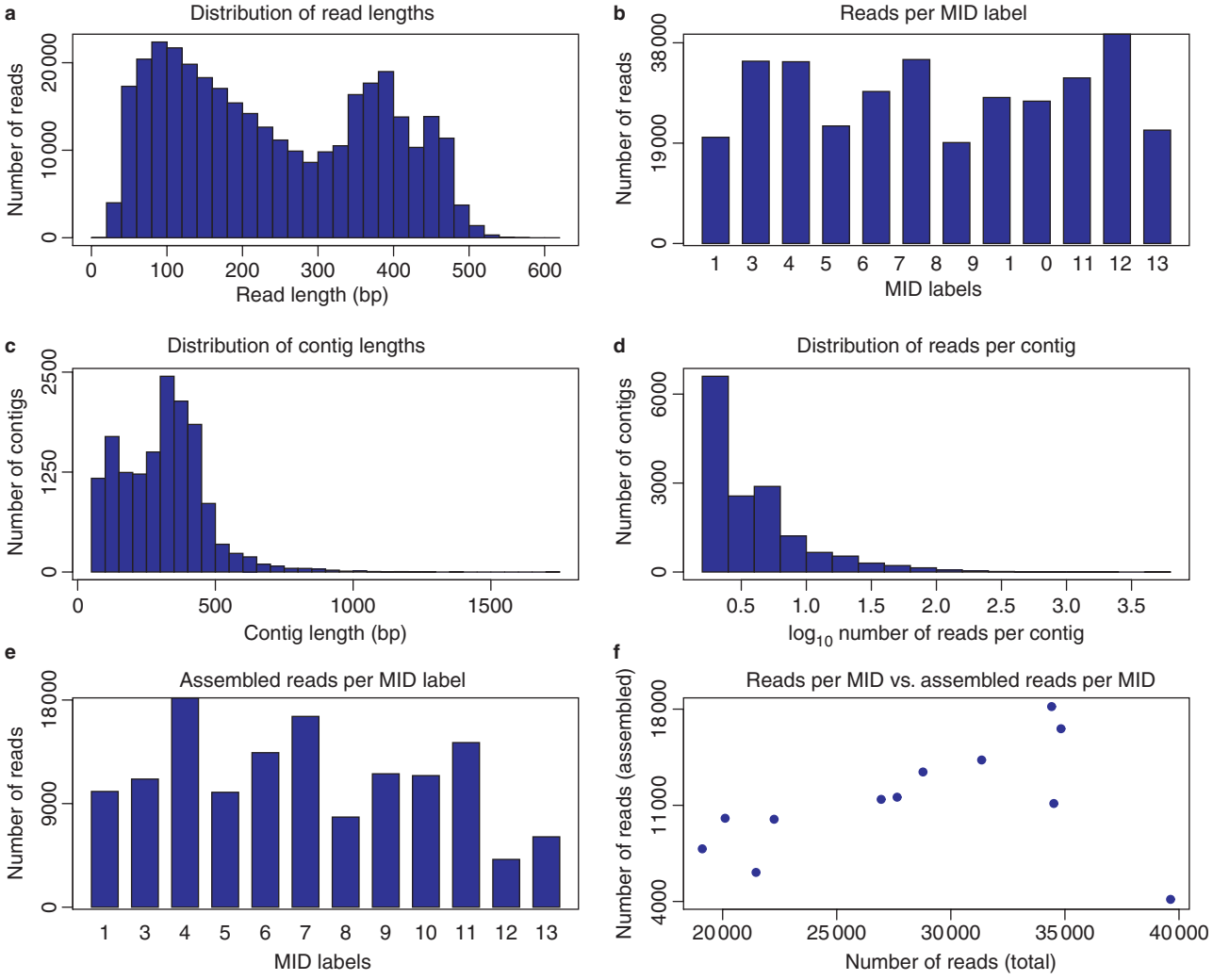

Fig. 2 Summary plots for 454 sequence reads and contigs. Plots show the distribution of read lengths in number of base pairs (a), the total number of reads obtained for each MID barcode (b), the distribution of contig lengths in number of base pairs (c), the distribution of the number of reads per contig (d), the number of reads that assembled into contigs for each MID barcode (e) and a scatter plot depicting the relationship between the number of reads obtained for each MID barcode and the number of reads assembled into contigs for each MID barcode (f).

from *Wolbachia* for the contigs and unassembled reads respectively. An additional 19 gene accessions from the contigs and 40 from the unassembled reads were characterized as being sequences from 'cellular organisms' and might also represent *Wolbachia* sequence data or other bacterial contaminants. This finding is consistent with previous studies suggesting that many *Lycaeides* populations are infected with *Wolbachia* (Gompert *et al.* 2008b; Nice *et al.* 2009).

BLAST search results suggest that a sizable portion of our 454 sequence data might represent sequence from genome regions associated with active or inactive transposable elements (including both retrotransposons and DNA transposons). Of the 828 unique UniRef50 gene accessions with significant similarity to our assembled contigs, approximately 100 relate to transposons (12.1%, rough approximation calculated by identifying genes with 'retro', 'integrase', 'reverse', 'gag', 'pol' or 'endonuclease' in their description). A similar portion of UniRef50 accessions with hits from the unassembled reads relate to transposons (167 of 1512, 11.0%). The frequency of BLAST hits to UniRef50 sequences relating to transposons is even higher if one counts all best BLAST hits not just unique best hits (contigs: 31.0%; unassembled reads: 15.7%).

Three hundred and twenty-one of the 828 unique UniRef50 accessions with significant similarity to our contigs could be associated with GO classifications. Top-level annotations were 154 accessions associated with cellular components, 303 with molecular functions

and 245 with biological processes (these annotations are not mutually exclusive). Six hundred and two of the 1512 unique UniRef50 accessions with significant similarity to the unassembled *Lycaeides* sequence reads were associated with GO classifications. These included 298 accessions associated with cellular components, 572 associated with molecular functions and 456 associated with biological processes. The distribution of GO classifications did not differ significantly between the contigs and unassembled reads (Fisher's exact test: P=0.9536).

Population genetic analyses

Including insertion–deletion polymorphisms, the mean proportion of variable sites within a contig was 0.0253 (SD 0.0377). Most contigs included at least one variable site; however, 2624 contigs were invariant (17.2%). As expected, the mean proportion of variant sites within a contig was lower when insertion–deletion polymorphisms were excluded (mean 0.0165, SD 0.0306) and nearly twice as many contigs were invariant (4547 contigs). We detected 57 776 insertion–deletion polymorphisms, which were sometimes, but not always, associated with homopolymer regions (number of flanking bases matching the base present at the insertion–deletion polymorphism: 0 for 20 076 sites, 1 for 10 000 sites, 2 for 7610 sites and ≥3 for 20 089 sites). As expected and regardless of the inclusion of insertion–deletion polymorphisms, we detected a moderate positive correlation between the number of reads in a contig and the proportion of sites that varied (with insertions–deletions: r=0.444, $P<2.2\times10^{-16}$; without insertion–deletions: r=0.475, $P<2.2\times10^{-16}$). The mean number of SNPs detected per population (MID), which does not include sites with insertion–deletion polymorphisms, was 1944.7 (SD 1086.3; Fig. S1). There was substantial among-population variation in the number of SNPs detected, which was highly correlated with the number of assembled reads for each population (r=0.8722, P= 0.000216). The estimated population frequency of the major allele (π_1) at each SNP varied from 0.5 to 1 (mean 0.8343, *sd* 0.1098). In general, the distribution of major allele frequencies suggests that substantial within-population genetic variation exists in the sampled populations. Estimates of the probability of sequence errors (ε) were generally low (mean 0.00499, SD 0.0220).

Our primary interest in Bayesian AMOVA was to estimate the posterior probability distribution for parameters of interest, as opposed to obtaining point estimates of those parameters. Nonetheless, point estimates and posterior credible intervals (CIs) provide convenient summaries of the parameters' posterior distributions, and are reported below. The median of the posterior probability distribution for genome-level ϕ_{ST} (μ_ϕ) was 0.3598 (95% CI 0.3371–0.3819), suggesting that approximately 36% of molecular variation was partitioned among *Lycaeides* populations (Fig. 3a). We detected substantial among-locus variation in estimates of ϕ_{ST} (Fig. 3b and c). Specifically, the average deviation between locus-specific estimates of ϕ_{ST} and genome-level ϕ_{ST} (σ_ϕ) was 0.3987 (95% CI 0.3650–0.4275) and posterior probability distributions for locus-specific ϕ_{ST} were not completely concordant. Estimates of ϕ_{ST} for a subset of highly variable loci are shown in Fig. S2.

The median and 95% CI from the posterior probability distributions of all pairwise genome-level ϕ_{ST} are given in Table 3. Median pairwise genome-level ϕ_{ST} ranged from 0.1224 (MID10 × MID12) to 0.3635 (MID5 × MID11). The former comparison included a *L. melissa* population (hypothesized central glacial

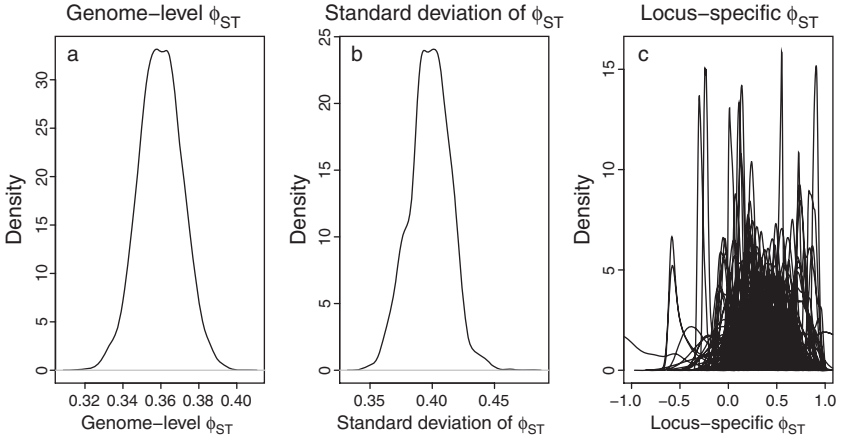

Fig. 3 Posterior probability distributions for genome-level ϕ_{ST} (a), the average deviation of locus-specific ϕ_{ST} from genome-level ϕ_{ST} (b) and locus-specific ϕ_{ST} for an arbitrary set of 200 loci (c). Posterior probability distributions were estimated from 24 000 MCMC iterations and were smoothed using a Gaussian kernel density function.

Table 3 Median (lower triangle) and 95% CI (upper triangle) of the posterior probability distributions for pairwise genome-level ϕ_{ST}

	MID1	MID3	MID4	MID5	MID6	MID7	MID8	MID9	MID10	MID11	MID12	MID13
MID1	0	0.2312–0.341	0.1168–0.2405	0.1878–0.2902	0.2468–0.3411	0.2316–0.3185	0.2719–0.3828	0.2879–0.3823	0.2457–0.346	0.3067–0.4094	0.1883–0.3594	0.1612–0.2961
MID3	0.2873	0	0.1403–0.2723	0.222–0.3318	0.1911–0.296	0.1368–0.2287	0.2358–0.3525	0.2259–0.3308	0.1722–0.2796	0.3102–0.4047	0.0854–0.2715	0.133–0.2697
MID4	0.1791	0.2057	0	0.0629–0.2015	0.1017–0.2119	0.1177–0.226	0.1749–0.317	0.1817–0.3087	0.1255–0.2462	0.2507–0.3753	0.0294–0.2474	0.1323–0.2875
MID5	0.2401	0.2772	0.134	0	0.1758–0.2749	0.1809–0.2714	0.202–0.3168	0.2366–0.3356	0.2021–0.301	0.3172–0.4104	0.0824–0.2523	0.1308–0.2625
MID6	0.2938	0.2441	0.1567	0.2253	0	0.1845–0.2676	0.1895–0.287	0.2213–0.3126	0.1799–0.2737	0.2742–0.364	0.0661–0.224	0.1888–0.3003
MID7	0.2751	0.1823	0.1729	0.2256	0.2273	0	0.1962–0.2891	0.2199–0.303	0.1785–0.2656	0.2793–0.3637	0.0756–0.2207	0.1394–0.241
MID8	0.327	0.2945	0.2461	0.2591	0.2393	0.242	0	0.1199–0.222	0.1507–0.2545	0.2863–0.3866	0.1186–0.3128	0.1372–0.274
MID9	0.335	0.2786	0.247	0.2866	0.2678	0.2631	0.171	0	0.1447–0.2381	0.2677–0.3572	0.1446–0.3005	0.1344–0.2557
MID10	0.2958	0.2254	0.1849	0.2511	0.2259	0.2221	0.2033	0.1913	0	0.2468–0.3342	0.0413–0.2028	0.0959–0.2162
MID11	0.3589	0.3579	0.3132	0.3635	0.3183	0.3221	0.3365	0.3127	0.29	0	0.1673–0.3201	0.2337–0.344
MID12	0.2754	0.1805	0.1392	0.1677	0.1461	0.1483	0.2179	0.2217	0.1224	0.2424	0	0.0162–0.2367
MID13	0.2294	0.2016	0.2099	0.1968	0.2454	0.1902	0.2053	0.1969	0.1561	0.2897	0.1262	0

refugium) and a putatively admixed *L. idas* population with ancestors from the hypothesized western and central refugia, whereas the latter comparison included a Karner blue (*L. melissa samuelis*) population (hypothesized eastern glacial refugium) and a population of the unnamed alpine hybrid species (hypothesized western and central refugial ancestors). In general, pairwise genome-level ϕ_{ST} suggest that molecular differentiation between the Karner blue population (MID11) and all other populations was particularly high (Table 3).

We were able to capture much of the structure described by pairwise genome-level ϕ_{ST} using NMDS with three dimensions (Fig. S3). Nonetheless, the ordination results from NMDS should be viewed as a simplification for graphical visualization, particularly as they are based solely on median values from the posterior probability distributions. Taken together, dimensions one and two separate populations hypothesized to be derived from ancestral western, central and eastern glacial refugial populations (Fig. 4). Specifically, the eastern refugium population (MID11) is separated from the western (MID1, MID3, MID4 and MID13) and central populations (MID8, MID9 and MID10) along dimension one, whereas central populations are separated from eastern and western populations along dimension two. Both dimensions one and two separate the putatively admixed populations (MID5, MID6, MID7 and MID12) from the eastern population, but these two dimensions do not clearly separate the putatively admixed populations from central or western populations (their hypothesized ancestors). Dimension three did not separate populations of any of the previously hypothesized glacial refugia (Fig. 4).

Discussion

Using 454 pyrosequencing, we obtained 341 045 sequence reads from 12 populations that we were able to assemble into 15 262 contigs representing one of the largest population genetic data sets for a non-model organism to date. The majority of these contigs were variable (82.8% including insertion–deletion polymorphisms or 70.2% including only SNPs) and thus potentially informative for estimating population genetic parameters. Although we used only 1570 of these contigs for the current analyses, additional contigs might be useful for subsets of these populations in future work. Sequence data from these contigs was sufficient to obtain precise estimate of overall and pairwise genome-level ϕ-statistics. The 454 sequence data showed evidence of population genetic structure in North American *Lycaeides*, as has been detected in previous studies of this genus (Nice *et al.* 2005; Gompert *et al.* 2006b, 2008a,b). Patterns of population genetic structure

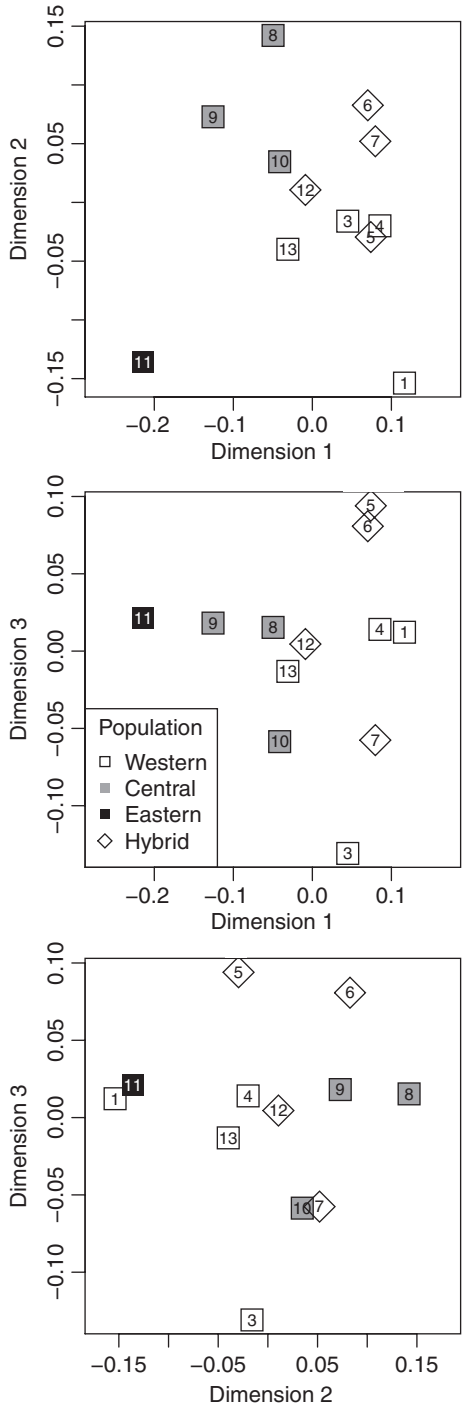

Fig. 4 Scatter plots depicting the ordination-based distances from NMDS among populations for pairs of dimensions. Populations are labelled with their MID barcode. Symbols denote population classifications according to hypothesized glacial refugia: western (grey square), central (white square), eastern (black square) or western × central admixed populations (white diamonds). Together dimensions one and two separate the western, central and eastern refugial populations, with putatively admixed populations occupying space in the vicinity of western and central populations.

based on genome-level pairwise ϕ_{ST}, inferred from our 454 data using a hierarchical Bayesian model for AMOVA, generally supported previously hypothesized glacial refugia and regions of postglacial secondary contact and hybridization (Nice *et al.* 2005; Gompert *et al.* 2006a,b, 2008b; Lucas *et al.* 2008) but also suggested that admixture may be more widespread than previously thought. However, it is important to note that even with precise multilocus estimates of population structure, it was not possible for us to exclude other potential hypotheses regarding the phylogeographic history of *Lycaeides*. Below we provide a detailed discussion of: (i) the phylogeographic history of *Lycaeides* in the light of these 454 data and (ii) the molecular methods we used, including considerations for population genetic studies using 454 sequencing.

Genetic variation, structure and phylogeographic history of Lycaeides

The proportion of variable sites in the *Lycaeides* 454 sequence data was on the order of that detected for other sequence markers examined in this genus (Gompert *et al.* 2006a, 2008a). However, approximately 35% of variable sites were insertion–deletion polymorphisms, which is a much higher percentage than was detected for other sequence markers in *Lycaeides* (including non-coding markers). The frequent occurrence of insertion–deletion polymorphisms might partially be a result of errors during pyrosequencing, particularly as 454 sequencing has difficulties with homopolymer regions (Margulies *et al.* 2005). However, many insertion–deletion polymorphisms were not associated with homopolymer regions. This potential source of error was partially responsible for our exclusion of insertion–deletion polymorphisms from SNP allele frequency estimation and Bayesian AMOVA.

Whereas the distribution of major allele frequencies for each of the 12 populations was relatively similar, the number of SNPs detected differed among populations by more than an order of magnitude. Our data suggest that this variation was primarily driven by variation in the number of assembled reads per population and not inherent differences in the levels of segregating variation within populations. We base this claim on the strong positive correlation detected between the number of SNPs found within a population and the number of assembled reads for a population (r=0.8722). Nonetheless, real differences in segregating genetic variation might exist among these populations, but addressing this question would require accounting for differences in coverage depth among populations and was outside of the scope of the current manuscript. Whether within-population

genetic diversity varies among the sampled populations could be determined by contrasting population-specific estimates of $\theta=4N_e\mu$, where N_e is the effective population size and μ is the mutation rate (e.g. Nei & Tajima 1981; Kuhner *et al.* 1995; Bernatchez & Wilson 1998).

Bayesian AMOVA that included all 12 sampled populations detected significant population genetic structure, as the 95% CI of the posterior distribution of the genome-level estimate of ϕ_{ST} (μ_ϕ) did not overlap zero. The posterior probability distribution of genome-level ϕ_{ST} indicated that between approximately 32% and 40% of molecular genetic variation was partitioned among populations. This is consistent with a phylogeographic history that included substantial periods of isolation among populations. Of course, this current level of population molecular differentiation probably reflects past and current geographic isolation, as well as some degree of reproductive isolation (Fordyce *et al.* 2002; Nice *et al.* 2002; Gompert *et al.* 2006a). The extent to which molecular genetic variation was partitioned among populations differed among loci, as evidenced by our estimate of σ_ϕ and the locus-specific estimates of ϕ_{ST}. This variation highlights the degree of stochasticity in the evolutionary process as captured by molecular data (Hein *et al.* 2005), and further emphasizes the importance of utilizing a large number of markers when estimating population genetic parameters (Funk & Omland 2003; Gompert *et al.* 2006b; Forister *et al.* 2008).

Pairwise genome-level ϕ_{ST} indicate that the Karner blue population (MID11) is clearly differentiated at the molecular level from all other populations. This is consistent with the Karner blue butterfly persisting in a distinct refugium during Pleistocene glacial cycles, as previously hypothesized (Nice *et al.* 2005). This level of differentiation cannot be easily explained by current geographic isolation or strong reproductive isolation, as the Karner blue population (MID11) is currently in close proximity to *L. melissa* populations (particularly MID10), with which it has hybridized leading to mitochondrial introgression (Gompert *et al.* 2006b, 2008b). Although our analyses also differentiate populations of the hypothesized western refugium (*L. idas*; MID1, MID3, MID4 and MID13) from populations of the hypothesized central refugium (*L. melissa*; MID8, MID9 and MID10), this separation is not nearly as clear. In several cases the degree of genetic differentiation between putative western and central populations (e.g. MID10 × MID13) is exceeded by that for pairs of western (e.g. MID1 × MID3) or central populations (e.g. MID8 × MID10).

When considered as a whole, putatively admixed populations are generally more genetically similar to hypothesized western and central refugial populations than these different refugial populations are to each other. This is consistent with the hypothesis of a history of hybridization in these populations. However, specific admixed populations are not clearly differentiated from specific western and central refugial populations (e.g. western refugial populations MID3 and MID4 cluster with hybrid populations MID5 and MID12). There are several possible explanations for this pattern. First, it is possible that some of the populations we classified as 'pure' western or central were in fact admixed. For example, previous data suggested that *L. idas* populations from northern California (MID3 and MID4) were descended from the hypothesized western refugium population but had experienced mitochondrial introgression from central populations (Gompert *et al.* 2008b). These populations showed little evidence of nuclear gene flow (hence our current classification). However, these previous results were based on AFLP markers, which might not have provided sufficient resolution to detect nuclear introgression. Thus, these northern California *L. idas* populations may in fact be admixed (this possibility is supported by other unpublished AFLP data). Another possibility is that post-Pleistocene range expansion was followed by geographic fragmentation and genetic differentiation among populations radiating from the same refugium. Given this alternative phylogeographic model it would be possible for admixed populations to be more similar to specific 'pure' populations than to others. This would result in a clustering in which some 'pure' and admixed populations are more similar than some 'pure' populations are to each other. It is also possible that these 'pure' populations that cluster close to hybrid populations may represent parental populations that contributed to the admixture. Finally, it is possible that the hypothesized western and central refugia did not exist but instead represent a single refugium. Thus, the western, central and putatively admixed populations would represent populations radiating from a single refugium at various points in time that have since become molecularly differentiated to varying degrees. However, this last scenario would be hard to reconcile with the sharp geographic transitions observed between differentiated populations and morphological intermediacy of putatively admixed populations. Moreover, this last hypothesis is not supported by previously published coalescent-based estimates of population divergence times or deep structure in mitochondrial data (Gompert *et al.* 2006a, 2008a). Evaluating these alternative models will require estimates of population divergence times and migration rates as well as detailed analyses of admixture. Future work using the 454 data presented here, as well as additional individual-based genetic data, will probably provide additional insights into these questions.

Lycaeides *454 data and future considerations*

A 1/2 PicoTiterPlate run on the 454 GS XLR70 Titanium platform provided us with sufficient data (341 045 reads) for precise estimation of population genetic parameters. Whereas many of these reads assembled into contigs (134 772 reads), many more could not be assembled into contigs (206 273 reads). Unassembled reads provide no information on the distribution of genetic variation, and thus are of little use for population genetic analyses using 454 sequence data. Therefore, the utility of our data for population genetic inference would have been improved had we further reduced the complexity of the *Lycaeides* DNA template by using more selective AFLP primers. Unfortunately, it is difficult to determine *a priori* the appropriate level of selectivity for AFLP primers, as many unique fragments will invariably have similar lengths and fragments from the same genetic region often differ in length. This problem could be ameliorated by conducting initial small-scale 454 pyrosequencing runs with varying degrees of genomic complexity reduction (i.e. different numbers of selective bases on AFLP primers) to determine the degree of selectivity that maximizes the number of reads that can be assembled while providing a sufficient total number of contigs (i.e. unique markers).

An additional potential problem with the AFLP-based technique we used is that it could bias population genetic inference. Specifically, variation is expected to exist within the restriction sites of AFLPs (hence their use as molecular markers), and only those reads with the presence of an AFLP restriction site will be sequenced. The degree that this will bias results should be related to the proportion of polymorphisms that occur within the restriction and priming sites of AFLP markers relative to the proportion that occur in the sequenced region between primer sites, as well as the overall degree of divergence among haplotypes. The EcoRI and MseI-C primers we used require specific 6- and 5-bp sequences for fragments to be generated from template DNA. Assuming polymorphisms are uniformly distributed across the genome, for a 310-bp AFLP fragment (the mean from our study) approximately 3–4% of polymorphisms should occur within restriction or priming sites leading to presence–absence variation of PCR products. Thus, most variation should be within the sequenced region and the potential bias introduced by certain haplotypes being excluded from amplification should be minimal. Nonetheless, this potential bias will increase with decreasing fragment size and should not be ignored. Finally, the probability that a substitution occurs within a restriction site will increase as the total number of substitutions among haplotypes increases. This will result in a decreased probability of obtaining sequence data for all haplotypes for genes that are highly variable and could lead to underestimates of genetic structure.

The AFLP-based technique we used for genomic complexity reduction is one of several options for obtaining data useful for population genetic inference from 454 pyrosequencing (or other next-generation sequencing technologies). Alternatively, specific PCR products from a large number of gene regions could be amplified independently and pooled for pyrosequencing (e.g. Bundock *et al.* 2009). This alternative procedure would allow for more precise control over the template used for pyrosequencing but would also require prior genomic knowledge for the design of primers. Additionally, cDNA could be used as a template for 454 pyrosequencing. 454 pyrosequencing from cDNA is common (and necessary) for transcriptome characterization (e.g. Vera *et al.* 2008; Hahn *et al.* 2009; Kristiansson *et al.* 2009; Meyer *et al.* 2009; Pauchet *et al.* 2009) but has not, to our knowledge, been used for population genetics. Assuming the number of transcribed genes are in the order of the number of genes desired for population genetic analysis, using cDNA instead of genomic DNA could result in the desired level of genomic complexity reduction. Using cDNA template for population genetics, however, would require cDNA normalization, might result in a decrease in the proportion of variable sites and the number of markers evolving neutrally, and creates assembly problems arising from alternative splicing of mRNA (Vera *et al.* 2008).

The ability to assign 454 sequence reads to populations is necessary for population genetic analyses. We were able to assign sequence reads to populations because we labelled amplified fragments with population-specific MID barcodes prior to pooling population templates for library construction and pyrosequencing. Although this procedure was generally successful, it resulted in variation among populations in the number of reads produced and assembled. For example, approximately three times as many reads assembled for MID4, MID7 and MID11 as for MID12. Generally the number of assembled reads was proportional to the number of reads, but this was not always the case (e.g. the most reads were obtained for MID12 but the least assembled). Variation in the number of reads per MID could simply represent stochastic variation during pyrosequencing, or could be partially due to bias in the amplification of specific MID barcodes as was suggested by van Orsouw *et al.* (2007). Although many population genetic analyses can be conducted without knowledge of individual-specific genotypes, this is not

true for analyses of linkage disequilibrium, individual-specific admixture, paternity, etc. Even when knowledge of an individual's genotypes is not required for population genetic inference, these data should generally decrease uncertainty in parameter estimates by decreasing sampling error. Thus, future studies should consider labelling individuals with MID barcodes. Currently, 151 MID barcodes are recognized by GS FLX Titanium software (454 Life Sciences Corp. 2009), allowing for a reasonable number of individuals to be labelled uniquely and pooled for PCR. This number will probably increase in future and could be circumvented presently with custom bioinformatics tools.

A relatively large proportion of our contigs and unassembled reads had significant BLAST hits to gene clusters in the UniRef50 database (14.7% and 10.7% respectively), suggesting that many of our contigs might correspond to transcribed regions of the genome. Arguably this relatively high proportion of putatively transcribed genomic regions is interesting, as we used genomic DNA that was not purposefully enriched for coding regions. The proportion of reads with significant UniRef50 BLAST hits was greater for the assembled contigs than for unassembled reads. This difference is not surprising, as many reads failed to assemble because they contained repeat regions, which are generally expected to occur more frequently in non-coding regions of the genome (e.g. Arcot *et al.* 1995; Ramsay *et al.* 1999). Repeat regions that could lead to problems with assembly include short tandem repeats and regions that have been duplicated throughout the genome. The latter include transposable elements, which might represent a sizable proportion of our sequence data (and thus the *Lycaeides* genome) but interestingly might be more common in contigs than in unassembled reads. GO annotations suggested a similar distribution of processes associated with genes from the assembled contigs and unassembled reads.

Finally, it is worth mentioning that although most of our sequence reads were from *Lycaeides*, bacterial contaminant sequences (most notably from *Wolbachia*) were present as well. Presumably, bacterial sequences can be easily identified and excluded from population genetic analyses (as we have done here), as bacterial genomes generally have fewer non-coding DNA regions than multicellular organisms and many bacterial genomes have been characterized (e.g. Blattner *et al.* 1997; Cole *et al.* 1998; Wu *et al.* 2004; Lynch 2007). However, contamination of 454 sequence data by foreign DNA (whether bacterial or from other sources) should be carefully guarded against for 454 sequence data generated without the use of taxon-specific primers (unfortunately, this could also pose a serious problem for AFLP studies in general).

Conclusions

The novel molecular and analytical methods we used in this study allowed us to precisely estimate measures of population genetic differentiation for North American *Lycaeides*. These estimates provide additional support for components of the hypothesized phylogeographic history of *Lycaeides* and suggest the possibility of nuclear admixture populations not previously thought to be admixed. Our results highlight the potential for next-generation sequencing technologies to provide the data necessary for high-precision estimation of population genetic parameters by labelling population or individual samples with unique sequence identification tags. Molecular methods similar to those we used in this manuscript should provide researchers working on non-model systems the tools necessary to generate large sequence data sets with little to no prior knowledge of the organism's genome. The need for such multilocus sequence data sets is clearly evidenced by the substantial variation observed in this study among locus-specific estimates of ϕ_{ST}. Finally, Bayesian methods in general, and hierarchical Bayesian models in particular, are likely to be very useful for analysing next-generation sequence data given the ability of these models to appropriately model uncertainty due to missing data and the hierarchical relationship between individual loci and the genome. The general hierarchical Bayesian framework we used for estimating ϕ_{ST} (i.e. modelling locus-specific parameters as random draws from a genome-level parameter distribution) can be used for estimating other population genetic parameters, including parameters from coalescent models (e.g. Storz & Beaumont 2002). Moreover, this model provides a framework for detecting genetic regions with ϕ_{ST} estimates that have low probability given the genome level ϕ_{ST} parameters and, thus, could serve as an outlier analysis for detecting selection (similar to the model proposed by Guo *et al.* 2009). We are currently pursuing this possibility. Hierarchical Bayesian models have become more prominent in population genetics (e.g. Storz & Beaumont 2002; Zhang *et al.* 2006; Guo *et al.* 2009) and will probably play increasingly greater roles in population genetics in future.

Acknowledgements

We thank the following for their comments on earlier versions of this manuscript: M. Brock, C. Edwards, L. Lucas, T. Parchman, two anonymous reviewers and H. Ellegren (*Molecular Ecology* subject editor). We are also indebted to M. Amaral, N. Anthony, R. ffrench-Constant, S. Fuller, G. Gelembiuk, L. Lucas and C. Schmidt for their assistance in the field. Thanks to the US Fish and Wildlife Service (USFWS permit PRT842392) and the US National Park Service (USNPS permits YELL-2008-SCI-5682

and GRTE-2008-SCI-0024) for granting us permission to collect *Lycaeides* specimens. We are particularly grateful to K. Geist for computational assistance with BLAST searches and Perl scripts. This work was funded by an NSF graduate research fellowship to ZG, a Professional Development Award from the University of Tennessee and NSF DEB 0614223 to JAF, a Research Enhancement Grant from Texas State University to CCN and NSF DBI 0701757 to CAB. MLF was supported by the Biology Department at the University of Nevada, Reno, and RJW was supported by an NIH INBRE program at the University of Wyoming for training in bioinformatics for undergraduates.

References

454 Life Sciences Corp. (2009) Using multiplex identifier (MID) adaptors for the GS FLX Titanium chemistry-extended MID set. Tech. rep., Technical Bulletin: Genome Sequencer FLX System.

Arcot SS, Wang ZY, Weber JL, Deininger PL, Batzer MA (1995) ALU repeats – a source for the genesis of primate microsatellites. *Genomics*, **29**, 136–144.

Avise JC (2004) *Molecular Markers, Natural History, and Evolution*. Chapman & Hall, London.

Bairoch A, Consortium U, Bougueleret L *et al.* (2009) The universal protein resource (UniProt) 2009. *Nucleic Acids Research*, **37**, D169–D174.

Barrett RDH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255–257.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Bentley DR (2006) Whole-genome re-sequencing. *Current Opinion in Genetics and Development*, **16**, 545–552.

Bernatchez L, Wilson CC (1998) Comparative phylogeography of nearctic and palearctic fishes. *Molecular Ecology*, **7**, 431–452.

Blattner FR, Plunkett G, Bloch CA *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Brookes MI, Graneau YA, King P, Rose OC, Thomas CD, Mallet JLB (1997) Genetic analysis of founder bottlenecks in the rare British butterfly *Plebejus argus*. *Conservation Biology*, **11**, 648–661.

Bundock PC, Eliott FG, Ablett G *et al.* (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal*, **7**, 347–354.

Cole ST, Brosch R, Parkhill J *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.

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.

Degnan JH, Rosenberg NA (2006) Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, 762–768.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Felsenstein J (1982) How can we infer geography and history from gene-frequencies? *Journal of Theoretical Biology*, **96**, 9–20.

Felsenstein J (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Fordyce JA, Nice CC, Forister ML, Shapiro AM (2002) The significance of wing pattern diversity in the Lycaenidae: mate discrimination by two recently diverged species. *Journal of Evolutionary Biology*, **15**, 871–879.

Forister ML, Nice CC, Fordyce JA, Gompert Z, Shapiro AM (2008) Considering evolutionary processes in the use of single-locus genetic data for conservation, with examples from the Lepidoptera. *Journal of Insect Conservation*, **12**, 37–51.

Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of native American mtDNA variation: a reappraisal. *American Journal of Human Genetics*, **59**, 935–945.

Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology Evolution and Systematics*, **34**, 397–423.

Galassi M, Davies J, Theiler J *et al.* (2009) *GNU Scientific Library: Reference Manual*. Network Theory Ltd, UK.

Gelman A, Carlin J, Stern H, Rubin D (2004) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall, London.

Gillespie J (2004) *Populations Genetics : A Concise Guide*, 2nd edn. Johns Hopkins University Press, Baltimore, Maryland.

Gnirke A, Melnikov A, Maguire J *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, **27**, 182–189.

Gompert Z, Fordyce JA, Forister ML, Shapiro AM, Nice CC (2006a) Homoploid hybrid speciation in an extreme habitat. *Science*, **314**, 1923–1925.

Gompert Z, Nice CC, Fordyce JA, Forister ML, Shapiro AM (2006b) Identifying units for conservation using molecular systematics: the cautionary tale of the Karner blue butterfly. *Molecular Ecology*, **15**, 1759–1768.

Gompert Z, Fordyce JA, Forister ML, Nice CC (2008a) Recent colonization and radiation of North American *Lycaeides* (*Plebejus*) inferred from mtDNA. *Molecular Phylogenetics and Evolution*, **48**, 481–490.

Gompert Z, Forister ML, Fordyce JA, Nice CC (2008b) Widespread mito-nuclear discordance with evidence for introgressive hybridization and selective sweeps in Lycaeides. *Molecular Ecology*, **17**, 5231–5244.

Guo F, Dey DK, Holsinger KE (2009) A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *Journal of the American Statistical Association*, **104**, 142–154.

Guppy C, Shepard J (2001) *Butterflies of British Columbia*. UBC Press, Vancouver.

Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL (2009) Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics*, **10**, 234.

Hale MC, McCormick CR, Jackson JR, DeWoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics*, **10**, 203.

Harismendy O, Ng PC, Strausberg RL *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**, R32.

Hedrick P (2005) *Genetics of Populations*, 3rd edn. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Hein J, Schierup M, Wiuf C (2005) *Gene Genealogies, Variation and Evolution*. Oxford University Press, Oxford.

Hellmann I, Mang Y, Gu Z *et al.* (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Research*, **18**, 1020–1029.

Hoarau G, Coyer JA, Veldsink JH, Stam WT, Olsen JL (2007) Glacial refugia and recolonization pathways in the brown seaweed *Fucus serratus*. *Molecular Ecology*, **16**, 3606–3616.

Hodges E, Rooks M, Xuan Z *et al.* (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols*, **4**, 960–974.

Holsinger KE, Weir BS (2009) Fundamental concepts in genetics: genetics in geographically structured populations: defining estimating and interpreting F_{ST}. *Nature Reviews Genetics*, **10**, 639–650.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Kristiansson E, Asker N, Forlin L, Larsson DGJ (2009) Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. *BMC Genomics*, **10**, 345.

Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population-size and mutation-rate from sequence data using Metropolis–Hastings sampling. *Genetics*, **140**, 1421–1430.

Kulathinal RJ, Stevison LS, Noor MAF (2009) The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genetics*, **5**, e1000550.

Lee JY, Edwards SV (2008) Divergence across Australia's Carpentarian barrier: statistical phylogeography of the Red-Backed Fairy Wren (*Malurus melanocephalus*). *Evolution*, **62**, 3117–3134.

Lessa EP (1990) Multidimensional-analysis of geographic genetic-structure. *Systematic Zoology*, **39**, 242–252.

Lucas LK, Fordyce JA, Nice CC (2008) Patterns of genitalic morphology around suture zones in North American *Lycaeides* (Lepidoptera: Lycaenidae): implications for taxonomy and historical biogeography. *Annals of the Entomological Society of America*, **101**, 172–180.

Lynch M (2007) *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, Massachusetts.

Lynch M (2008) Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Molecular Biology and Evolution*, **25**, 2409–2419.

Lynch M (2009) Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics*, **182**, 295–301.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Meyer E, Aglyamova GV, Wang S *et al.* (2009) Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GS FLX. *BMC Genomics*, **10**, 219.

Nabokov V (1949) The nearctic members of *Lycaeides* Hubner (Lycaenidae, Lepidoptera). *Bulletin of the Museum of Comparative Zoology*, **101**, 479–541.

Nei M, Tajima F (1981) DNA polymorphism detectable by restriction 910 endonucleases. *Genetics*, **97**, 145–163.

Nice CC, Shapiro AM (1999) Molecular and morphological divergence in the butterfly genus *Lycaeides* (Lepidoptera: Lycaenidae) in North America: evidence of recent speciation. *Journal of Evolutionary Biology*, **12**, 936–950.

Nice CC, Fordyce JA, Shapiro AM, Ffrench-Constant R (2002) Lack of evidence for reproductive isolation among ecologically specialised lycaenid butterflies. *Ecological Entomology*, **27**, 702–712.

Nice CC, Anthony N, Gelembiuk G, Raterman D, ffrench-Constant R (2005) The history and geography of diversification within the butterfly genus *Lycaeides* in North America. *Molecular Ecology*, **14**, 1741–1754.

Nice CC, Gompert Z, Forister ML, Fordyce JA (2009) An unseen foe in arthropod conservation efforts: the case of *Wolbachia* infections in the Karner Blue butterfly. *Biological Conservation*, **14**, 3137–3146.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Novaes E, Drost DR, Farmerie WG *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 312.

Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007) Microarray-based genomic selection for high-throughput resequencing. *Nature Methods*, **4**, 907–909.

Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH (2008) Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. *Bioinformatics*, **24**, 2776–2777.

van Orsouw NJ, Hogers RCJ, Janssen A *et al.* (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One*, **2**, e1172.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Patin E, Laval G, Barreiro LB *et al.* (2009) Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genetics*, **5**, e1000448.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, **441**, 1103–1108.

Pauchet Y, Wilkinson P, van Munster M, Augustin S, Pauron D, ffrench Constant RH (2009) Pyrosequencing of the midgut transcriptome of the poplar leaf beetle *Chrysomela tremulae* reveals new gene families in Coleoptera. *Insect Biochemistry and Molecular Biology*, **39**, 403–413.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, ISBN 3-900051-07-0.

Ramsay L, Macaulay M, Cardle L *et al.* (1999) Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. *Plant Journal* , **17**, 415–425.

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the co-ancestry coefficient: basis for a short-term genetic-distance. *Genetics*, **105**, 767–779.

Scott J (1986) *The Butterflies of North America: A Natural History and Field Guide*. Stanford University Press, Stanford, California.

Slatkin M (1987) Gene flow and the geographic structure of natural-populations. *Science*, **236**, 787–792.

Slatkin M (1993) Isolation by distance in equilibrium and nonequilibrium populations. *Evolution*, **47**, 264–279.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457–462.

Storz JF, Beaumont MA (2002) Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution*, **56**, 154–166.

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.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.

Taberlet P, Fumagalli L, Wust-Saucy AG, Cosson JF (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molecular Ecology*, **7**, 453–464.

Venables WN, Ripley B (2002) *Modern Applied Statistics with S*, 4th edn. Springer Verlag, New York.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 0097–0159.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Wu M, Sun LV, Vamathevan J *et al.* (2004) Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. *PLoS Biology*, **2**, 327–341.

Zhang Y, Niu T, Liu JS (2006) A coalescence-guided hierarchical Bayesian method for haplotype inference. *American Journal of Human Genetics*, **79**, 313–322.

Zach Gompert is a PhD student in the ecology program at the University of Wyoming. His research interests include speciation genetics, hybridization, and Bayesian methods in population genetics. Matt Forister, assistant professor at the University of Nevada, Reno, has research interests that include diet specialization and the evolutionary ecology of plant-insect interactions. James Fordyce is an associate professor at the University of Tennessee with research interests in ecological factors that promote population differentiation and maintain variation. Chris Nice is an associate professor at Texas State University with interests in evolutionary ecology and genetics. Robert Williamson is an undergraduate at the Rose-Hulman Institute of Technology with interests in sexual selection and computational biology. Alex Buerkle is an associate professor at the Univeristy of Wyoming with interests in evolutionary genetics, hybridization, and speciation.

Supporting Information

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

Fig. S1 Boxplots depicting the distribution of major allele frequencies for each population (MID). The number of SNPs per population with 5× coverage are given above each boxplot. Major allele frequencies were estimated by maximum likelihood while accounting for sequence errors.

Fig. S2 Posterior probability distributions for locus-specific φ_{ST} for 125 loci with more ≥25 haplotypes (a) and six loci with ≥90 haplotypes (b). Posterior probability distributions were estimated from 24 000 MCMC iterations and were smoothed using a Gaussian kernel density function. These loci posses levels of variation that are similar to levels of variation for loci generally used in population genetic analyses.

Fig. S3 Scree plot (a) and Shepard plot (b) from NMDS analysis. The former shows the relationship between the number of dimensions used for NMDS and the sum of squared differences between the ordination-based distances and the distances predicted by regression (stress). The latter depicts the relationship between genome-level pairwise φ_{ST} and ordination-based distances (points) as well as the predicted values from regressing the latter on the former (solid line). These plots suggest that the three dimensions used for NMDS captured the structure of the pairwise φ_{ST} well.

Fig. S4 Boxplots depicting the distribution of major allele frequencies for simulated data. Dashed blue lines denote the major allele frequency used for each set of simulations.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.