Electronic Supplementary Material

Pyrosequencing and sampling localities

The data analyzed in the present paper comes from pyrosequencing of *Lycaeides* populations as described in detail in Gompert *et al.* (2010). The "populations" involved in pyrosequencing each consisted of 15 individuals from three sampling locations, the details of which are given in supplemental table 1. Individuals within populations were pooled to create population samples that were processed using restriction enzymes in a complexity-reduction step (van Orsouw *et al.* 2007; Baird *et al.* 2008) and labeled with unique sequence barcodes prior to being sequenced on a 454 GS XLR70 Titanium platform by SeqWright DNA Technology Services (Houston, TX, USA). Subsequent to pyrosequencing, barcodes were used to associate individual sequence reads with populations. Barcodes were then removed from the sequences and the SeqMan NGen assembler v.2.0.0 (DNASTAR) was used to assemble contigs; further details on the parameterization of assembly are provided in Gompert *et al.* (2010), and the data are available from the NCBI short read archive (accession number SRA010351). Custom Perl scripts were used to select contigs for analyses in IM using the constraint (as discussed in the main text) of at least 3 sequences per population per contig for each pairwise comparison between populations.

Implementation of the Isolation with Migration Model

Here we discuss additional details in the implementation of the IM model using the program IMa, and discuss assumptions of the model. Initial runs of IMa were conducted to find effective parameters for the MCMC and to set priors that constrained searches, yet included the full range of reasonable parameter space. We used the geometric increment model for 10 chains, with 0.8 specifying the non-linearity of "heating" across chains, and 0.9 specifying the heating level for the highest numbered chain. Trend lines and effective sample sizes were inspected to verify chain mixing and convergence of parameter estimates. Mutations models implemented in IMa for sequence data (as opposed to microsatellites) are the infinite sites model and HKY (Strasburg & Rieseberg 2010). The infinite sites model was not appropriate for our

data given the possibility of recurrent mutation and/or recombination, as suggested by the results of the four gamete test (Hudson & Kaplan 1985) reported in supplemental table 2. The fraction of contigs that failed the four gamete test ranged from 13 to 26%. These failures could be caused by either recurrent mutation or recombination, but in any event are not apparently biased by population (and thus should not introduce a bias to analyses). We note in particular that the variation among sets of contigs used for the same population in different comparisons is similar to the variation across populations (e.g. *L. melissa* east in the comparison with KBB had a failure rate of 14%, while 26% of contigs failed in the comparison with *L. melissa* west). Other details involving molecular variation are shown in supplemental table 2.

As discussed in the main text (Materials & Methods), we do not have a genome-wide estimate for mutation rate (μ) in *Lycaeides*. Therefore, we have not converted the parameters estimated by IMa into values of years or individuals. Rather, values are left in the units estimated by the model: for example, migration (*m*) is a ratio of m / μ (where m is the rate per gene per generation), and the population size estimates (Θ) are 4 x Ne (effective population size) x μ. Note that μ is per gene (genetic region or locus) and is a geometric mean across loci, thus variation in fragment size could be important. However, fragment sizes are not biased in any way across our populations (the average number of SNPs per contig is quite similar across populations, as shown in supplemental table 2).

The IM model as implemented in IMa makes a number of assumptions, including that the genetic regions being studied are not physically linked and are not under selection. Our samples were pooled at the population level, thus we did not test for linkage disequilibrium. If any of our markers are linked, we assume that the large number of markers compensates. Regarding the assumption of neutral variation, we do not, at present, have a means for estimating the extent to which the genomes in question are under selection (though the majority of our data come from non-coding regions (Gompert *et al.* 2010)). Outlier analysis (Beaumont & Nichols 1996; Foll & Gaggiotti 2008) is not appropriate given the average number of sequences in our contigs (see main text for details). However, this is not a limitation for our use of the IM model, as we are interested in genetically-effective rates of gene flow in a comparative sense. If selection acts against admixed individuals, and thereby reduces their contribution to a gene pool, this will

be reflected in lower rates of gene flow as estimated by the IM model. Because our data were generated by pyrosequencing of pooled groups of individuals, it is possible that the data includes multiple sequences of the same gene copy for a given individual. This is a violation of the IM model as these replicate sequences should coalesce immediately. However, this problem should be minimal as we have relatively few sequences compared to the number of sampled gene copies in each DNA pool.

Finally, the IM model assumes a bifurcating mode of diversification in which the populations being analyzed are each other's closest relatives, and are not exchanging genes with other populations. The history of *Lycaeides* is complex, and still being explored, but sister relationships for KBB and *L. melissa*, and for *L. melissa* and *L. idas* are reasonable given the current state of knowledge (Nice *et al.* 2005, Gompert *et al*. 2008). Our pairwise comparisons likely violate the assumption of no gene flow with other populations. This assumption is commonly violated in studies that employ the IM model, as few natural pairs of populations will have no connections to other populations. However, simulations have suggested that IM is robust to modest violations of most of the underlying assumptions (Strasburg & Rieseberg 2010).

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. & Johnson, E. A. 2008 Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One* **3**.

Beaumont, M. A. & Nichols, R. A. 1996 Evaluating loci for use in the genetic analysis of population structure. Proceedings of the Royal Society of London Series B-Biological Sciences **263**, 1619-1626.

Foll, M. & Gaggiotti, O. 2008 A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. Genetics **180**, 977-993.

Gompert, Z., Forister, M. L., Fordyce, J. A. & Nice, C. C. 2008 Widespread mito-nuclear discordance with evidence for introgressive hybridization and selective sweeps in *Lycaeides*. *Molecular Ecology* **17**, 5231-5244.

Gompert, Z., Forister, M. L., Fordyce, J. A., Nice, C. C., Williamson, R. J. & Buerkle, C. A. 2010 Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology* **19**, 2455-2473.

Hudson, R. R., & Kaplan N. L. 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147-164.

Nice, C. C., 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.

Strasburg, J. L. & Rieseberg, L. H. 2010 How Robust Are "Isolation with Migration" Analyses to Violations of the IM Model? A Simulation Study. *Molecular Biology and Evolution* **27**, 297-310.

Suzek, B. E., Huang, H. Z., McGarvey, P., Mazumder, R. & Wu, C. H. 2007 UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288.

van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J., Verstegen, H. & van Eijk, M. J. T. 2007 Complexity Reduction of Polymorphic Sequences (CRoPS (TM)): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. *Plos One* **2**.

Supplemental table 1. Locality information for populations studied.

Taxon	Locality	Latitude	Longitude
KBB	Fort McCoy, WI	43°47’59’’N	90°49’59’’W
	Necedah, WI	44°04’00’’N	90°11’20’’W
	Saratoga, NY	43°03’24’’N	73°48’47’’W
L. melissa east	Brandon, SD	43°36’29’’N	96°34’39’’W
	Victor, ID	43°39’32’’N	111°06’41’’W
	Indian Bathtubs, WY	41°12’14’’N	106°46’17’’W
L. melissa west	Garderville, NV	38°48’54’’N	119°46’44’’W
	Verdi, NV	39°03’01’’N	119°55’48’’W
	Sierra Valley, CA	39°37’48’’N	120°21’40’’W
L. idas	Yuba Gap, CA	39°29’34’’N	120°35’39’’W
	Leek Springs, CA	38°37’59’’N	120°14’24’’W
	Trap Creek, CA	39°22’43’’N	120°40’27’’W

125

Supplemental table 2. Details of data used in analyses. "Comparison" here refers to the pairs of populations involved in IMa analyses, as in table 1. Results from the four gamete test are shown in the "Number failed out of total" column, with number referring to the number of contigs.

Comparison	Total base pairs	Population	Number failed out of total (percent)	SNPs	Proportion variable sites	Average SNPs per contig (SE)	GC content
KBB and *L. melissa* east	91944	KBB	73 / 317 (23%)	2521	0.0274	7.95 (0.48)	49.7%
		L. melissa east	45 / 317 (14%)	2233	0.0243	7.04 (0.45)	48.3%
L. melissa west and *L. idas*	60362	*L. melissa* west	25 / 188 (13%)	1229	0.0204	6.54 (0.52)	47.2%
		L. idas	44 / 188 (23%)	1621	0.0269	8.62 (0.63)	47.5%
L. melissa west and *L. melissa* east	71031	*L. melissa* west	36 / 236 (15%)	1512	0.0213	6.41 (0.45)	47.9%
		L. melissa east	62 / 236 (26%)	2100	0.0296	8.90 (0.64)	50.0%