MOLECULAR ECOLOGY

Molecular Ecology (2010) **19**, 2361–2363

NEWS AND VIEWS

PERSPECTIVE

Next generation population genetics and phylogeography

KENT E. HOLSINGER

Department of Ecology & Evolutionary Biology, U-3043, University of Connecticut, Storrs, CT 06269-3043, USA

Received 2 April 2010; revision received 12 April 2010; accepted 16 April 2010

In March, 2010 *Molecular Ecology* published a special issue on 'Next generation molecular ecology'. The papers published in that issue covered topics ranging from high-throughput sequencing of environmental DNA to genome-wide SNP detection and analysis of alternative splicing to comparative transcriptome sequencing. Clearly new sequencing technologies will 'result in transformation of how we think about molecular ecology as a discipline', as Tautz *et al.* (2010) argued in their introduction to that special issue. We see another example in this issue, where Gompert *et al.* (2010) use 454 pyrosequencing to gain new insight into the complex phylogeographic history of *Lycaeides* butterflies.

Earlier work in *Lycaeides* relied either on sequence data from just a few genomic regions, e.g. SSCP analysis of 236 bp from the AT-rich region of the mitochondrial genome (Nice *et al.* 2005) and DNA sequence analysis of 1000 bp from COI and COII (Gompert *et al.* 2008a) or on a modest number of fragment-based markers (Gompert *et al.* 2006a, 2008b). In contrast, Gompert *et al.* (2010) now identify nearly 1600 contigs with greater than 10 times coverage in all of the populations they study. Since the average length of these contigs is about ~250 bp, the results they present are based on roughly 400 kb of DNA sequence [Gompert *et al.* (2010) report a mean contig length of 310 bp, but only contigs <700 bp were used in population comparison to avoid repetitive regions that may belong to non-orthologous genes]—two orders of magnitude greater than was available only 2 years ago. The result is very precise estimates both of the overall amount of genetic differentiation among populations and of the pairwise divergences among the sampled populations.

Gompert *et al.* (2010) use Φ_{ST} (Excoffier *et al.* 1992) to measure genetic differentiation among populations and find that 36% of variation is associated with among-population differences (95% CI: 0.34–0.38). Φ_{ST} can also be interpreted as a measure of the extent to which isolation among populations increases the coalescence time for genes drawn randomly from different populations relative to the coalescence time for genes within a single population (Slatkin 1991; Holsinger & Weir 2009). Thus, pairwise estimates of Φ_{ST} between populations provide insight into the degree to which populations are historically connected, although they cannot reveal the extent to which such connections reflect contemporary gene flow vs. recent common ancestry of populations that are now isolated (Felsenstein 1982).

In *Lycaeides*, a non-metric multidimensional scaling analysis of pairwise Φ_{ST} estimates reveals three broad geographic groupings (Fig. 1): (i) a western group corresponding roughly to what has been referred to as *L. idas*; (ii) a central group corresponding roughly to *L. melissa*; and (iii) an eastern group corresponding to *L. melissa samuelis* (the Karner blue). Estimates of Φ_{ST} from pairwise comparisons involving the Karner blue were larger than those involving other populations, generally greater than 0.3 and greater than 0.24 in every case. In contrast, the western and central groups of populations are more weakly differentiated. Some pairs of western populations and some

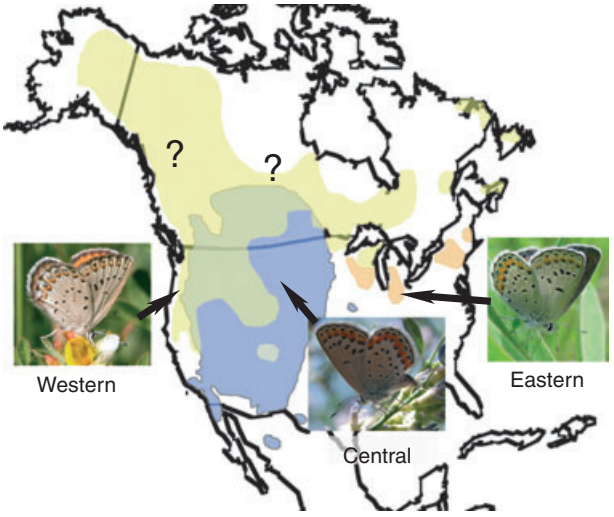

Fig. 1 A stylized summary of the phylogeography of North American *Lycaeides* butterflies. The approximate current distribution of individuals derived from each glacial distribution is shown for the western (yellow), central (blue) and eastern (orange) refugia. The region of overlap and hybridization between the western and central lineages is shown in green. Some Canadian populations might be derived from a fourth refugium (denoted with question marks). Female *Lycaeides* from each lineage are shown on their larval host plant: *Lotus nevadensis* (western, photo by James Fordyce), *Medicago sativa* (central, photo by Lauren Lucas) and *Lupinus perennis* (eastern, photo by Dave Hanson). Individuals from the eastern refugium use the latter host plant exclusively.

Correspondence: Kent E. Holsinger, Fax: +1 (860) 486-6364; E-mail: kent@darwin.eeb.uconn.edu

pairs of central populations have greater pairwise Φ_{ST}s than do other pairs in which one population belongs to the western group and the other belongs to the central group.

The geographic groups correspond with hypothesized glacial refugia and it is tempting to interpret the smaller amount of differentiation between western and central populations as a result of gene exchange associated with recent secondary contact. Indeed, Gompert *et al.* (2006b, 2008b) point out that other data are consistent with this hypothesis. But central (*L. melissa*) and eastern (Karner blue) populations are geographically adjacent and mitochondrial introgression has been detected between the Karner blue and populations of *L. melissa* (Gompert *et al.* 2006b, 2008b). Thus, the larger pairwise differences for comparisons involving the Karner blue might simply reflect smaller population sizes in the Karner blue (Weir 1996; Hedrick 1999; Holsinger& Weir 2009). The Karner blue was listed as an endangered species under the US Endangered Species Act in 1992 and its populations are restricted to remnant savannas and barrens, primarily in Wisconsin and Michigan (Wooley 2003).

The mean estimates of population differentiation just summarized provide a great deal of insight, but they also mask many differences among loci. Using a hierarchical Bayesian model similar to the ones introduced by Beaumont & Balding (2004) and Guo *et al.* (2009), Gompert *et al.* (2010) estimate that the standard deviation of Φ_{ST} across loci was 0.40 (95% CI: 0.37–0.43), and locus-specific estimates of Φ_{ST} vary from as little as –0.5 to nearly 1.0 (Gompert *et al.* 2010; Fig. 3C). This variation reflects real variation in the amount of differentiation at each locus, not statistical uncertainty associated with the genome-wide estimate of Φ_{ST}. That real variation undoubtedly reflects to some degree the enormous variability inherent in the process of genetic drift, but some of it is also likely to reflect the effects of natural selection at loci subject to different patterns of natural selection. Those effects may either be associated with loci included in the sample or with loci that are closely linked. One of the advantages of the Bayesian model Gompert *et al.* (2010) use is that it can be extended to identify statistical 'outliers' where the amount of among-population differentiation may reflect the effects of natural selection (Beaumont & Balding 2004; Guo *et al.* 2009).

In addition to illustrating how next generation sequencing may transform studies of population genetics and phylogeography, the *Lycaeides* data also illustrate some of the challenges that lie ahead. Take for example the 'simple' question: which of the populations included in this sample is the most genetically diverse? Gompert *et al.* (2010) point out that the proportion of variable sites detected in contigs is positively correlated with the number of reads ($r = 0.44$ with indels, $r = 0.48$ without indels). Thus, simply calculating the number of variable sites or expected heterozygosity in the sample is not enough. Investigators wishing to compare levels of diversity among populations will either have to use rarefaction methods like those in (Mousadik & Petit 1996) or they will have to estimate $\theta = 4N_e\mu$ using an approach like the composite likelihood method introduced in Hellmann *et al.* (2008). Moreover, as Hellman *et al.* (2008) illustrate, with whole-genome shotgun sequencing it becomes possible to compare levels of diversity in different parts of the genome. In organisms lacking a reference genome, like *Lycaeides*, such intragenomic comparisons will necessarily be limited to comparing levels of diversity among isolated contigs. Nonetheless, such comparisons will provide another way in which to identify regions of the genome subject to different patterns of natural selection.

Nearly 40 years ago the field of evolutionary genetics was transformed when Harris (1966), Hubby & Lewontin (1966) and Lewontin & Hubby (1966) introduced protein electrophoresis to the field. We stand on the threshold of a similar transformation today, and studies like the analysis of *Lycaeides* presented by Gompert *et al.* (2010) give us a glimpse of the insights that are sure to follow.

References

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

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.

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.

Gompert Z, Forister ML, Fordyce JA *et al.* (2010) Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology* **19**, 2455–2473.

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.

Harris H (1966) Enzyme polymorphisms in man. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **164**, 298–310.

Hedrick PW (1999) Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution*, **53**, 313.

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.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Review Genetics*, **10**, 639–650.

Hubby JL, Lewontin RC (1966) A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. *Genetics*, **54**, 577–594.

Lewontin RC, Hubby JL (1966) A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics*, **54**, 595–609.

Mousadik Ae, Petit RJ (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theoretical and Applied Genetics*, **92**, 832–839.

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

Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research*, **58**, 167–175.

Tautz D, Ellegren H, Weigel D (2010) Next generation molecular ecology. *Molecular Ecology*, **19**, 1–3.

Weir BS (1996) . *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*, Sinauer Associates, Sunderland, Massachussets.

Wooley CM (2003) Notice of availability of the approved recovery plan for the Karner Blue butterfly (*Lycaeides melissa samuelis*). *Federal Register*, **68**, 54913–54914.

doi: 10.1111/j.1365-294X.2010.04667.x