Molecular Phylogenetics and Evolution xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories ☆

Jeffrey M. DaCosta *, Michael D. Sorenson

Department of Biology, Boston University, Boston, USA

ARTICLE INFO

Article history:
Received 9 February 2015
Revised 22 July 2015
Accepted 29 July 2015
Available online xxxx

Keywords:
ddRAD-seq
Phylogenomics
Species tree methods
Lagonosticta
Vidua
Brood parasitism

ABSTRACT

Genotype-by-sequencing (GBS) methods have revolutionized the field of molecular ecology, but their application in molecular phylogenetics remains somewhat limited. In addition, most phylogenetic studies based on large GBS data sets have relied on analyses of concatenated data rather than species tree methods that explicitly account for genealogical stochasticity among loci. We explored the utility of "double-digest" restriction site-associated DNA sequencing (ddRAD-seq) for phylogenetic analyses of the *Lagonosticta* firefinches (family Estrildidae) and the *Vidua* brood parasitic finches (family Viduidae). As expected, the number of homologous loci shared among samples was negatively correlated with genetic distance due to the accumulation of restriction site polymorphisms. Nonetheless, for each genus, we obtained data sets of ~3000 loci shared in common among all samples, including a more distantly related outgroup taxon. For all samples combined, we obtained >1000 homologous loci despite ~20 my divergence between estrildid and parasitic finches. In addition to nucleotide polymorphisms, the ddRAD-seq data yielded large sets of indel and locus presence–absence polymorphisms, all of which had higher consistency indices than mtDNA sequence data in the context of concatenated parsimony analyses. Species tree methods, using individual gene trees or single nucleotide polymorphisms as input, generated results broadly consistent with analyses of concatenated data, particularly for *Lagonosticta*, which appears to have a well resolved, bifurcating history. Results for *Vidua* were also generally consistent across methods and data sets, although nodal support and results from different species tree methods were more variable. Lower gene tree congruence in *Vidua* is likely the result of its unique evolutionary history, which includes rapid speciation by host shift and occasional hybridization and introgression due to incomplete reproductive isolation. We conclude that ddRAD-seq is a cost-effective method for generating robust phylogenetic data sets, particularly for analyses of closely related species and genera.

1. Introduction

Molecular phylogenetics has evolved during the past decade due to conceptual, technological, and analytical/computational developments. Among these has been a growing appreciation for the stochasticity of lineage sorting, resulting in variation among gene trees and potential incongruence with the species trees (Pamilo and Nei, 1988; Page and Charleston, 1997; Edwards and Beerli, 2000; Nichols, 2001; Rosenberg and Nordborg, 2002). This has led to an increasingly important interface between population genetics and phylogenetics (Edwards, 2009), and a shift toward datasets comprising multiple nuclear loci that more broadly sample the stochastic sorting of ancestral polymorphisms and the accumulation of informative variation along the internodes separating speciation events. Multi-locus datasets have in turn spurred the development of new phylogenetic methods that accommodate genealogical stochasticity in the estimation of phylogeny (e.g., Edwards et al., 2007; Liu, 2008; Heled and Drummond, 2010; reviewed in Knowles and Kubatko, 2010).

At the same time, advances in DNA sequencing technology have led to the development of methods that sample hundreds to thousands of genomic loci in a rapid and cost-effective manner. For most applications in population genetics and systematics, however, harnessing the power of this new technology requires methods that consistently recover a set of homologous loci across

☆ This paper was edited by the Associate Editor Derek Wildman.

* Corresponding author at: Department of Biology, Boston University, 5 Cummington Mall, Boston, MA 02215, USA.

E-mail address: dacostaj@bu.edu (J.M. DaCosta).

http://dx.doi.org/10.1016/j.ympev.2015.07.026
1055-7903/© 2015 Elsevier Inc. All rights reserved.

multiple samples (Davey et al., 2011; McCormack et al., 2013). The number of loci targeted should be large enough to take full advantage of the current throughput of DNA sequencers but still small enough to allow for adequate coverage for a large set of multiplexed samples and loci processed in a single run, thereby minimizing cost. Methods based on multiplexing a large number of locus-specific primer pairs (e.g., Binladen et al., 2007) or using hybridization probes to isolate portions of the genome for shotgun sequencing (e.g., Mamanova et al., 2010) require prior knowledge of genomic sequences for the taxa of interest to design conserved primers/probes, an approach that has associated strengths and weaknesses (McCormack et al., 2013). An alternative approach is to use restriction-site associated DNA sequencing (RAD-seq) (Miller et al., 2007; Baird et al., 2008), which uses one or more restriction enzymes to target homologous loci among a set of samples and therefore requires no *a priori* genomic resources. The method provides a cost-effective means of broadly sampling the genome and, along with similar "genotype-by-sequencing" (GBS) methods, has emerged as a powerful and increasingly popular tool for an array of population level applications in molecular ecology (Davey and Blaxter, 2010; Narum et al., 2013).

Relatively few studies have explored the utility of RAD-seq data for phylogenetic analysis (e.g., Near and Benard, 2004; Rubin et al., 2012; Eaton and Ree, 2013; Jones et al., 2013; Keller et al., 2013; Nadeau et al., 2013; Wagner et al., 2013; Cruaud et al., 2014; Hipp et al., 2014). As with other methods based on restriction sites, an important complication of using RAD-seq in species level (or higher) phylogenetic analyses is the inevitable reduction in the number of homologous loci captured among samples as mutations in enzyme recognition sites accumulate with increasing genetic divergence among samples. Rubin et al. (2012) explored this issue by computationally extracting RAD loci from reference genomes of *Drosophila*, mammals, and fungi. While these simulated RAD-seq data produced accurate topologies for *Drosophila* and for shallow nodes in the mammalian and fungal trees, nodes representing speciation events >60 Mya were not reliably recovered due in part to a dearth of homologous RAD loci recovered from highly divergent taxa. Patterns of locus recovery among taxa, reflecting the gain and loss of enzyme recognition sites, represent a potentially useful source of characters for phylogenetic analysis. This approach, however, has not been tested; all RAD-seq phylogenetic studies to date have relied on nucleotide character matrices.

Applying methods that accommodate incomplete lineage sorting in the estimation of species trees to large RAD-seq datasets also presents a computational challenge. To date, empirical RAD-seq phylogenetic studies have largely side-stepped this issue by using concatenated data and assuming that phylogenetic signal in the aggregate overrides noise generated by loci that conflict with the underlying species tree (e.g., Nadeau et al., 2013; Wagner et al., 2013; Cruaud et al., 2014; but see Eaton and Ree, 2013). This approach may fail, however, if any relationships within the species tree fall within the "anomaly zone," in which adding data leads to stronger confidence in a tree that is discordant with the species tree (Degnan and Rosenberg, 2006; Kubatko and Degnan, 2007; Rosenberg and Tao, 2008; but see Huang and Knowles, 2009). In addition to the computational load of processing a large number of loci, the application of more sophisticated methods may be hindered by the difficulty of estimating individual gene trees and/or model parameters from the typically short sequences generated by RAD-seq. For example, Bayesian methods (Liu, 2008; Heled and Drummond, 2010) that simultaneously estimate a large set of parameters from thousands of short loci are unlikely to converge on reliable parameter estimates even when using a simplified model (e.g., linking model and rate parameters over many loci). While each locus may provide relatively little phylogenetic information, the volume of loci generated by RAD-seq produces a large number of single nucleotide polymorphisms (SNPs), and new methods that use SNP data to infer species trees (Bryant et al., 2012; Kubatko and Chifman, 2014) may be more effective.

We explored the utility of RAD-seq data for phylogenetic analysis using a variety of character matrices and tree inference methods, and compared results for two avian genera with contrasting ecology and evolutionary histories. The firefinches (genus *Lagonosticta*) comprise 10 species that are distributed in sub-Saharan Africa (Fry and Keith, 2004), and likely diversified over the course of several million years through a typical process of allopatric speciation (Sorenson et al., 2004). Firefinches are the primary hosts of the brood parasitic indigobirds, which together with whydahs comprise the genus *Vidua*, which has a total of 19 recognized species. Imprinting on hosts during development shapes the courtship and mating behavior of adult indigobirds (i.e., songs and mate choice preferences), providing a mechanism for rapid speciation when a novel host is colonized (Payne, 1973; Payne et al., 2000; Sorenson et al., 2003). The behaviors that drive speciation by host shift in indigobirds are shared by most other *Vidua* (i.e., the whydahs), appear to be ancestral in the clade, and may also lead to hybridization between established parasitic species (Payne and Sorenson, 2004). Thus, post-speciation gene flow in this clade may obscure species-level relationships. The contrasting evolutionary histories of these two genera are apparent in a comparison of mitochondrial DNA (mtDNA) phylogenies (Sorenson et al., 2004). Relationships among *Lagonosticta* species are generally well resolved, whereas the branching order of *Vidua* clades is uncertain. In this study, we present results of phylogenetic analyses for each genus based on thousands of "double-digest" RAD-seq (ddRAD-seq) loci (Peterson et al., 2012; DaCosta and Sorenson, 2014); we implement a variety of phylogenetic methods/approaches and test the utility of different categories of character information, including SNPs, indels, and the presence–absence of loci in the data set. We also evaluate the "decay" of homologous ddRAD-seq loci with increasing genetic divergence among samples and compare the extent of gene-tree/species-tree congruence between the two genera.

2. Materials and methods

2.1. Taxon sampling, DNA sequencing, and bioinformatics

Tissue samples were collected during fieldwork in Cameroon and Tanzania, or obtained from tissue collections at natural history museums (Table 1). The *Lagonosticta* dataset includes nine *Lagonosticta* samples representing seven species and a single brown twinspot (*Clytospiza monteiri*) sample as the outgroup; mtDNA sequence data support a sister group relationship between these two genera (Sorenson et al., 2004). The *Vidua* dataset includes 14 *Vidua* samples representing 12 species and a single cuckoo finch (*Anomalospiza imberbis*) sample as the outgroup (Sorenson and Payne, 2001). Genomic DNA was extracted from tissue samples using a DNeasy Tissue Kit (Qiagen Inc.). Mitochondrial DNA sequences from each taxon were collected following previously described methods (Sorenson and Payne, 2001; Sorenson et al., 2004). In some cases, different samples of a given species were used for mtDNA and ddRAD sequencing (see Table 1). We collected sequence data from the following mtDNA regions, comprising a total of 2186 base pairs (bp): tRNA methionine, NADH dehydrogenase subunit 2 (ND2), tRNA tryptophan, NADH dehydrogenase subunit 6 (ND6), tRNA glutamic acid, and control region. Mitochondrial sequences for each taxon were aligned and reconciled in Sequencher v4 (Gene Codes Corporation); interspecific alignments were manually edited in Se-Al v2 (http://tree.bio.ed.ac.uk/software/seal/).

Please cite this article in press as: DaCosta, J.M., Sorenson, M.D. ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories. Mol. Phylogenet. Evol. (2015), http://dx.doi.org/10.1016/j.ympev.2015.07.026