Information dropout patterns in restriction site associated DNA phylogenomics and a comparison with multilocus Sanger data in a species-rich moth genus

Lee, Kyung Min

2018-11

Lee , K M , Kivelä , S M , Ivanov , V , Hausmann , A , Kaila , L , Wahlberg , N & Mutanen , M 2018 , ' Information dropout patterns in restriction site associated DNA phylogenomics and a comparison with multilocus Sanger data in a species-rich moth genus ' , Systematic Biology , vol. 67 , no. 6 , pp. 925-939 . https://doi.org/10.1093/sysbio/syy029 , https://doi.org/10.1093/sysbio/syy029

http://hdl.handle.net/10138/307940
https://doi.org/10.1093/sysbio/syy029

acceptedVersion

Running head: DROPOUT PATTERNS IN RAD PHYLOGENOMICS

Title: Information Dropout Patterns in RAD Phylogenomics and a Comparison with Multilocus Sanger Data in a Species-rich Moth Genus

Authors:

Kyung Min Lee[1], Sami M. Kivelä[1,5], Vladislav Ivanov[1], Axel Hausmann[2], Lauri Kaila[3], Niklas Wahlberg[4] & Marko Mutanen[1*]

Authors' affiliations:

[1] *Department of Ecology and Genetics, University of Oulu, Finland*

[2] *SNSB – Bavarian State Collection of Zoology, Munich, Germany*

[3] *Finnish Museum of Natural History, Zoology Unit, University of Helsinki, Finland*

[4] *Department of Biology, Lund University, Sweden*

[5] *Current address: Department of Zoology, Institute of Ecology and Earth Sciences, University of Tartu, Vanemuise 46, EE-51014 Tartu, Estonia*

Authors' email addresses:

Kyung Min Lee: kyungmin.lee@oulu.fi

Sami M. Kivelä: sami.mikael.kivela@ut.ee

Vladislav Ivanov: vladislav.ivanov@oulu.fi

Axel Hausmann: axel.hausmann@zsm.mwn.de

Lauri Kaila: lauri.kaila@helsinki.fi

Niklas Wahlberg: niklas.wahlberg@biol.lu.se

Marko Mutanen: marko.mutanen@oulu.fi

Correspondence author address, fax number and e-mail (*):

*Marko Mutanen

University of Oulu

Department of Ecology and Genetics

P.O. Box 3000

FI-90014 University of Oulu

Finland

Tel: +358 (0)8 553 1256

Fax: +358 (0)8 344 064

Email: marko.mutanen@oulu.fi

Abstract. A rapid shift from traditional Sanger sequencing-based molecular methods to the phylogenomic approach with large numbers of loci is underway. Among phylogenomic methods, RAD (Restriction site Associated DNA) sequencing approaches have gained much attention as they enable rapid generation of up to thousands of loci randomly scattered across the genome and are suitable for non-model species. RAD data sets however suffer from large amounts of missing data and rapid locus dropout with decreasing relatedness among taxa. The relationship between locus dropout and the amount of phylogenetic information retained in the data has remained largely un-investigated. Similarly, phylogenetic hypotheses based on RAD have rarely been compared with phylogenetic hypotheses based on multilocus Sanger sequencing, even less so using exactly the same species and specimens. We compared the Sanger-based phylogenetic hypothesis (8 loci; 6,172 bp) of 32 species of the diverse moth genus *Eupithecia* (Lepidoptera, Geometridae) to that based on double-digest RAD sequencing (3,256 loci; 726,658 bp). We observed that topologies were largely congruent, with some notable exceptions that we discuss. The locus dropout effect was strong, making our data set a borderline case for RAD approach in terms of phylogenetic resolution at deepest phylogenetic levels. We demonstrate that locus number is not a precise measure of phylogenetic information content of the data since, due to the short time for mutations to have accumulated, the number of single-nucleotide polymorphisms (SNPs) may remain low at very shallow phylogenetic levels despite large numbers of loci. As we hypothesize, the number of SNPs and parsimony informative SNPs (PIS) first increase towards deeper phylogenetic levels even if the associated effects of increased hierarchical redundancy are eliminated, the number of SNPs peaking at intermediate phylogenetic levels and, thereafter, declining again as a result of decay of available loci. Similarly, we indicate with empirical data that the locus dropout affects the type of loci retained, the loci found in many species tending to show lower interspecific distances than those shared among fewer species. We also examine the effects of the numbers of loci, SNPs and PIS on nodal bootstrap support, but could not demonstrate with our data our expectation of a positive

Commented [KLJ1]: along?

Commented [KLJ2]: This is either no more discussed in the text, or I don't find it discussed.

correlation between them. We conclude that RAD methods provide a useful tool for phylogenomics as indicated by its broad congruence with an eight-gene Sanger data set if study organisms are not too closely or too distantly related to each other. Focus should be on distribution and number of SNPs and PIS rather than on loci available for phylogenetic inference at different phylogenetic depths. **Key words:** Allelic dropout, ddRAD sequencing, *Eupithecia*, Lepidoptera, Locus dropout, Molecular systematics, Parsimony informative SNPs, RAD sequencing, SNP dropout

Commented [KLJ3]: this is how I had originally understood the punch line

High-throughput DNA sequencing methods have enabled rapid generation of genome-wide DNA sequence data simultaneously from many specimens with reasonable costs. Several NGS sequencing platforms have become available (Mardis 2013) and a number of different methods have been developed to accumulate data to address specific scientific questions, including various areas of systematic research (Lemmon and Lemmon 2013). Recent approaches include anchored hybrid enrichment (Lemmon et al. 2012; Brandley et al. 2015; Hamilton et al. 2016; Breinholt et al. 2017) and several varieties of restriction site associated DNA sequencing (RAD) (Miller et al. 2007; Baird et al. 2008). RAD methods, based on the digestion of genomic DNA with restriction enzymes and subsequent sequencing of short regions adjacent to the restriction sites, enable efficient SNP (single nucleotide polymorphism) discovery and are being used to infer phylogenetic relationships (Eaton and Ree 2013; Wang et al. 2013; Hipp et al. 2014; Hou et al. 2015) (Dasmahapatra et al. 2012; Nadeau et al. 2013; Jones et al. 2013; Keller et al. 2013; Cruaud et al. 2014; Takahashi et al. 2014; Pante et al. 2015; Ebel et al. 2015; Gonen et al. 2015; Herrera et al. 2015; Leaché et al. 2015b; McCluskey and Postlethwait 2015; DaCosta and Sorenson 2016) (Leaché et al. 2014; Herrera and Shank 2016). The use of RAD tags has usually resulted in well-resolved phylogenies, although trials are not numerous, and only a few have been conducted on truly diverse groups.

Several RAD-based studies have focused on young species groups and taxonomically complex groups with horizontal gene transfer and incomplete lineage sorting potentially complicating inferring phylogenies or species trees (Eaton and Ree 2013; Rheindt et al. 2014; Streicher et al. 2014). Other studies have been carried out with well-defined and even arguably relatively old (ten to tens of millions years) species (Rubin et al. 2012; Cruaud et al. 2014; Hipp et al. 2014; Viricel et al. 2014; Herrera et al. 2015; McCluskey and Postlethwait 2015; Herrera and Shank 2016; Eaton et al. 2017). Of the RAD methods, double-digest RAD sequencing (ddRADseq) has a benefit of high repeatability because it avoids the random shearing characteristic of traditional RAD methods, which makes combining independent datasets straightforward as long as the same restriction

Commented [MM4]: Kyung Min: combine these together and remove the excluded references

enzyme pair has been used (Peterson et al. 2012; Kai et al. 2014; Puritz et al. 2014). So far, only a few explorations of ddRADseq have been conducted in a phylogenetic context (Kai et al. 2014; Leaché et al. 2015a; DaCosta and Sorenson 2016).

RAD-based approaches have several benefits (Davey and Blaxter 2010; Rowe et al. 2011; Puritz et al. 2014). Restriction sites are scattered all over the genome and therefore RAD tags provide an overview of the entire genome. Typically, the analysis yields thousands of loci (ca. 100-150 bp fragments) and SNPs per specimen. Alcohol preserved specimens are suitable and since reads are relatively short (usually 50-150 bp), dry collection specimens with degraded DNA are potentially useful as well (Tin et al. 2014; Suchan et al. 2016). Furthermore, the efficient use of RAD tags does not require a reference genome. Therefore, the method is suitable for non-model organisms (Andrews et al. 2016; Kim et al. 2016).

Commented [KLJ5]: cf discussion where we state otherwise

In spite of these benefits, RAD sequencing has certain limitations. RAD tags typically consist of substantial amounts of missing data, potentially complicating the inference of phylogenetic relationships (Rubin et al. 2012; Lemmon and Lemmon 2013; Wagner et al. 2013; DaCosta and Sorenson 2016). Attention has been directed to recognizing orthologous loci and distinguishing them from paralogous loci (Rubin et al. 2012; Cariou et al. 2013; Gonen et al. 2015). Another major practical issue is that the likelihood to recover an orthologous locus is negatively correlated with time since the divergence of the lineages where the compared individuals belong to, because mutations are gradually accumulated on restriction sites as time elapses. Thus, only a fraction of shared loci is recovered between genetically distant individuals, arguably reducing the efficacy of the method at deeper phylogenetic levels (Arnold et al. 2013; Ree and Hipp 2015). Indeed, several studies have indicated that rapid locus dropout (also called locus decay or allelic dropout) is an inherent feature of RAD data and the effect can be drastic (Gonen et al. 2015; Leaché et al. 2015b; DaCosta and Sorenson 2016). If mutation rate remains constant in time, a linear dropout of loci is expected with decreasing relatedness between two lineages (Fig. 1). Loci recovered between distant

144 relatives are expected to be slowly evolving (e.g. protein coding genes), which translates into a
145 disproportionately low number of SNPs and consequently a weak phylogenetic signal, further
146 exaggerating the data decay at deep phylogenetic levels (Leaché et al. 2015a). Huang and Knowles
147 (2016) demonstrated with simulated data that low tolerance to missing data leads to a
148 disproportionately high exclusion rate of loci with high mutation rate. Locus dropout and decreased
149 mutation rate of retained loci are complementary and predict a constant steep loss of information
150 towards deeper phylogenetic levels. Eaton et al. (2017) recently demonstrated that, somewhat
151 counter-intuitively, the influence of locus dropout on the phylogenetic information content at deeper
152 phylogenetic levels is less significant than previously expected because decay of phylogenetic
153 information resulting from locus dropout is compensated by the increase of taxa towards the deeper
154 nodes. Consequently, Eaton et al. (2017) concluded that the negative effects of locus dropout can be
155 mitigated by increasing taxon sampling.

156 We recognize an additional effect inherent to RAD data sets, which differs from the previously
157 recognized effects in a remarkable way. Previous studies have largely concentrated on the sequence
158 data amount *per se*, but such measures do not provide a reliable picture of the amount of
159 phylogenetic information content in the data. This is because phylogenetic relatedness is highly
160 correlated with genetic similarity. Consequently, at very shallow phylogenetic levels, the number of
161 retrieved loci can be very high, while at the same time they may be poor in phylogenetic
162 information due to the limited time for mutations to have accumulated (Fig. 1). We therefore predict
163 that the number of SNPs and PIS decrease towards very shallow phylogenetic levels and peaks at
164 intermediate phylogenetic levels. As a result, the phylogenetic information content is not supposed
165 to be linearly correlated with the number of loci. In Figure 1, the expected relationship between the
166 loci and SNPs/PIS along with increasing coalescence time between two lineages is demonstrated in
167 a schematic way. To our best knowledge, the relationship between locus and SNP/PIS dropouts
168 across phylogenetic time has not been investigated.

Here, we aim at assessing the potential of ddRADseq in resolving phylogenetic affinities in the looper moth genus *Eupithecia* Curtis (vernacular name 'pugs') (Lepidoptera, Geometridae) and conduct a detailed examination of patterns and effects of loci, SNPs and PIS on ddRAD phylogeny. *Eupithecia* is one of the most diversely radiated metazoan genera and includes 1,362 described valid species world-wide (Scoble and Hausmann 2007). Species of *Eupithecia* show high levels of morphological similarity and niche specialization (McDunnough 1949; Mironov 2003), both features characterizing many megadiverse insect groups. Due to the high number of species and close morphological similarity, attempts to resolve their relationships with rigorous methodology are virtually lacking.

We start by examining effects of ddRAD locus parameters (clustering threshold and minimum number of individuals per locus) on ddRAD tree topology and confidence. We continue by examining the congruence between the eight-gene Sanger data set and the ddRAD phylogenies. Few similar comparisons have previously been carried out (but see Cruaud *et al.* 2014; Ruane et al. 2015). The Sanger phylogeny of *Eupithecia* is constructed based on a set of one mitochondrial and seven nuclear genes that combined have repeatedly shown to have high information value at intermediate and deep phylogenetic levels in Lepidoptera (e.g. Mutanen *et al.* 2010; Sihvonen *et al.* 2011; Zahiri *et al.* 2012; Heikkilä *et al.* 2015). We next examine how the level of locus conservativeness is related to SNP/PIS abundance and investigate if locus and SNP/PIS distributions at different phylogenetic depths follow the predicted patterns as presented in Figure 1. Finally, we statistically examine locus and SNP/PIS effects on nodal support values.

Commented [KLJ6]: There is stuff in material& methods that should be placed here.

MATERIAL AND METHODS

Taxon sampling

193 We sampled a total of 42 specimens from 35 species of *Eupithecia* that were collected during
194 2006-2014 from Finland, Germany and Italy. *Pasiphila rectangulata* was also included to serve as
195 outgroup, both genera belonging to the tribe Eupitheciini (Larentiinae). Detailed information on the
196 specimens' label data is provided in Table S1. All specimens subjected to DNA sequencing were
197 assigned a label with a unique sample ID. One or two legs of each specimen were deposited in
198 microplate wells, each filled with 30 µl of absolute ethanol.

199

200 *Molecular methods*

201 Sanger sequencing was performed for one mitochondrial and seven nuclear markers. The
202 sequencing for the mt COI gene was carried out at the Canadian Centre for DNA Barcoding
203 (CCDB) following laboratory protocols used routinely in CCDB as explained in detail in DeWaard
204 et al. (2008). In order to proceed with the sequencing for nuclear genes and the ddRAD library
205 preparation, genomic DNA (gDNA) was separately extracted from two legs using the DNeasy
206 Blood & Tissue Kit (Qiagen) in the molecular laboratory at the University of Oulu, Finland. All
207 PCR and sequencing protocols followed Wahlberg and Wheat (2008), except for PCR clean-up that
208 was carried out with ExoSAP-IT (Affymetrix) and Sephadex columns (Sigma-Aldrich) and
209 sequencing that was done using an ABI 3730 DNA Analyzer (Applied Biosystems). We collected
210 sequence data from the following nuclear regions comprising a total of 6,172 base pairs (bp):
211 carbamoylphosphate synthase domain protein (CAD), elongation factor 1 alpha (EF1α),
212 glyceraldhyde-3-phosphate dehydrogenase (GAPDH), isocitrate dehydrogenase (IDH), cytosolic
213 malate dehydrogenase (MDH), ribosomal protein S5 (RpS5), wingless (see Table S2). All
214 sequences for each taxon were manually aligned and edited using BioEdit (Hall 1999). All DNA
215 sequences are available at the U.S. National Center for Biotechnology Information (NCBI)
216 GenBank (Accessions xx – xx).

Double-digested RAD-Seq libraries were prepared following Peterson et al. (2012). All samples were whole-genome amplified prior to experimentation using a REPLI-g Mini kit (Qiagen) due to low concentrations of gDNA in the original isolates. Concentration of the amplified gDNA was estimated with the PicoGreen kit (Molecular Probes) according to the kit instructions. 200 ng of gDNA was digested with *Pst*I and *Mse*I restriction enzymes (New England Biolabs). Following digestion, ligation of double-stranded sequencing adapters was completed in the same tube. The P1 adapter included the Illumina sequencing primer sequences, one of 43 unique, five bp barcodes, and a TGCA overhang on the top strand to match the sticky end left by *Pst*I. The P2 adapter included the Illumina sequencing primer sequences and an AT overhang on the top strand to match the sticky end left by *Mse*I. It also incorporated a ‘‘divergent-Y’’ to prevent amplification of fragments with *Mse*I cut sites on both ends. Following ligation, the size selection was performed by automated size-selection technology, BluePippin (Sage Science; 2% agarose cartridge). We produced two pooled libraries in four lanes of the machine using automated size selection set to “tight” with a mean of 300 bp. Size selected libraries were eluted in 40 µL volumes and enriched by PCR using library-specific indexed primers complementary to the Illumina paired-end adapters. Amplified DNA fragments were purified with AMPure XP magnetic beads (Agencourt). The quality, size and concentration of the pooled libraries were finally determined using the MultiNA® (Shimadzu). Individual fragment libraries were then combined in equimolar amounts and sequenced on an Illumina HiSeq 2500 PE 100. DNA reads from ddRAD sequencing are available at the NCBI Sequence Read Archive (SRA) [BioProject ID: PRJNA345300]. To rule out contamination by the bacterial parasite *Wolbachia*, the ddRAD reads were mapped to *Wolbachia pipientis* (GenBank: NZ_JQAM01000001) using Geneious 10.0.9 (Biomatters).

ddRADseq data processing, examination of effects of locus parameters and assessing comprehensiveness of data

242 We processed raw Illumina reads using the pyRAD v.3.0.5 (Eaton 2014) pipeline. This program
243 is designed to assemble data for phylogenetic studies that contain divergent species using global
244 alignment clustering which may include indel variation. We de-multiplexed samples using their
245 unique barcode and adapter sequences, and sites with Phred quality scores below 20 were converted
246 to "N" characters, and reads with ≥ 10% N's were discarded. The filtered reads for each sample
247 were clustered using the program VSEARCH v.1.1.3 (VSEARCH GitHub repository,
248 https://github.com/torognes/vsearch), and then aligned with MUSCLE v.3.8.31 (Edgar 2004). This
249 clustering step establishes homology among reads within a species. As an additional filtering step,
250 such consensus sequences were discarded that had low coverage (< 3 reads), excessive
251 undetermined or heterozygous sites (> 10) potential resulting from paralogs or highly repetitive
252 genomic regions, or too many haplotypes (> 2 for diploids). In addition, we excluded loci with
253 excessive (> 3) shared polymorphic sites as likely representing clustering of paralogs. The
254 consensus sequences were clustered across samples at 80, 85, 90, 95% similarity. This step
255 establishes locus homology among species. The justification for this filtering method is that shared
256 heterozygous SNPs across species are more likely to represent a fixed difference among paralogs
257 than shared heterozygosity within orthologs among species. We applied a strict filter that allowed a
258 maximum of three species to share heterozygosity at a given site (paralog = 3).

259 The final ddRADseq loci were assembled by adjusting a minimum number of individuals per
260 locus (*m*) value, which specifies the minimum number of individuals that are required to have data
261 present at a locus in order for that locus to be included in the final matrix. Our ddRADseq dataset
262 contained 43 individuals from 36 species (35 *Eupithecia* species and *Pasiphila rectangulata* as
263 outgroup), and setting m=6 retains loci with data present for three or more species. By contrast,
264 setting m=43 retains zero loci with data present for all individuals (= 100% complete matrix). We
265 compiled data matrices with *m* values of each 4, 6, 9, 12, 15, 21 to determine the potential impact of

Field Code Changed

number of loci, SNPs, parsimony informative SNPs (PIS), and missing data on phylogenetic analysis.

We generated a pairwise similarity matrix for individuals based on locus-sharing patterns using RADami v. 1.0-3 (Hipp et al. 2014) in R 3.1.3 (R Core Team 2015). This analysis returned a pairwise similarity matrix based on how many loci or the proportion of loci shared between individuals. Proportions of locus-sharing across all specimens were plotted on a graph.

We assessed the comprehensiveness of our dataset by comparing the number and proportion of observed loci retained at the sequencing depth used in the final data sets ($d \geq 3$; d denotes the sequencing depth) with those of observed showing depth less than 3 (observed 1-3 times).

Construction of reference assembly data set

We also constructed a phylogenetic hypothesis based only on the reads that we could map on available lepidopteran genomes. For the reference assembly, we used the following 26 genomes as reference: *Amyelois transitella* [GCF_001186105], *Bombyx mori* [GCF_000151625], *Calycopis cecrops* [GCA_001625245], *Chilo suppressalis* [GCA_000636095], *Danaus plexippus* [GCA_000235995], *Heliconius cydno*, [GCA_001485745] *H. elevatus* [GCA_900068365], *H. ethilla*, [GCA_001485985] *H. hecale* [GCA_001486065], *H. ismenius* [GCA_001485965], *H. melpomene* [GCA_000313835], *H. numata* [GCA_900068715], *H. pardalinus* [GCA_001486225], *H. timareta* [GCA_001486185], *Lerema accius* [GCA_001278395], *Manduca sexta* [GCA_000262585], *Melitaea cinxia* [GCA_000716385], *Operophtera brumata* [GCA_001266575], *Papilio glaucus* [GCA_000931545], *Papilio machaon* [GCF_001298355], *Papilio polytes* [GCF_000836215], *Papilio xuthus* [GCF_000836235], *Phoebis sennae* [GCA_001586405], *Pieris rapae* [GCA_001856805], *Plutella xylostella* [GCF_000330985], and *Spodoptera frugiperda* [GCA_002213285]. We concatenated these genomes to a single reference file. Sequences were

assembled using *ipyrad* v.0.7.11 (Eaton and Overcast 2016). Reads were trimmed of barcodes and adapters and quality filtered using a q-score threshold of 33, with bases below this score converted to Ns and any reads with more than 5 Ns removed. Reads were mapped to the concatenated reference genomes with *BWA* based on sequence similarity using the default *bwa mem* setting. With the collected reads, similar clusters of reads were identified using a threshold of 85% of similarity and were aligned. Next, we performed joint estimation of heterozygosity and error rate based on a diploid model assuming a maximum of 2 consensus alleles per individual. We then used the parameters from the previous step, heterozygosity and error rate, to determine consensus base calls for each allele, and removed consensus sequences with greater than 5 Ns per end of paired-end reads. With consensus sequences identified, step six clustered and aligned reads for each sample to consensus sequences. Finally, we filtered the dataset according to maximum number of indels allowed per read end (8), maximum number of SNPs per locus (20), maximum proportion of shared heterozygous sites per locus (0.5), and minimum number of samples per locus (3).

Construction of phylogenetic trees

To infer a phylogenetic hypothesis, we used concatenated sequences from all recovered RAD loci. We used the maximum likelihood (ML) method implemented in the RAxML 8.2.0 (Stamatakis 2006) program with a GTR+GAMMA model for nucleotide substitutions for phylogeny constructions. Two hundred independent trees were inferred, applying options of automatically optimized subtree pruning regrafting (SPR) rearrangement and 25 distinct rate categories in the program to identify the best tree. Statistical support for each branch was obtained using the rapid algorithm from 500 bootstrap replicates under the same substitution model.

For reference assembly data, the ML tree was built using the unpartitioned GTR+CAT model and branch support was assessed by a 500 replicates rapid-bootstrap analysis. The following species

were not included in the reference assembly due to the low number of recovered loci: *E. tantillaria*, *E. tenuiata*, *E. linariata*, *E. intricata*, *E. nanata*, *E. centaureata*, *E. vulgata* and *E. abietaria*.

Effects of locus conservativeness on SNP frequency

We investigated whether locus conservativeness is correlated with SNPs in our data, ande expected that conservative loci are shared more widely among individuals a. Number of SNPs/locus supposedly decreases as the number of individuals/locus increases. We fitted generalized linear models with a negative binomial error distribution and logarithmic link function (R function 'glm.nb' [Venables and Ripley 2002]) to the data derived with $m \geq 6$, lower values of m being excluded due to the risk of alien loci (e.g. of bacterial origin) to be included in the data. To assess potential non-linearity of the relationship between the number of SNPs/locus and the number of individuals/locus, we compared models where the linear predictor included only a linear term for the number of individuals/locus and a model with both the linear and quadratic terms. Models were compared based on their AIC and BIC values. Because the normal distribution assumption of residuals was violated in both models, we further derived 95% adjusted bootstrap percentile confidence intervals for the mean number of SNPs/locus with each value of m (individuals/locus), excluding the cases where less than seven observations were available ($m \geq 21$). Bootstrap analyses (10,000 resamples) were conducted with the R functions 'boot' and 'boot.ci' (Canty and Ripley 2015).

Commented [KLJ7]: I think that this further explanation of the null hypothesis is not needed to be repeated here. Is the previous sentence needed either?.

Commented [KLJ8]: Actually, is anything above at correct place here, in methods, or needed at all? If needed, in introduction unless already there. The same applies some other 'introductionary' chapters of various methods below as well.

Patterns of locus, SNP and PIS dropout and their effects on nodal confidence

We used node depth as a proxy for node age (in relative terms) and used nodes as observation units. In order to quantify the depth values for each node, we converted the ML tree into an ultrametric tree (Fig. S1) based on rate smoothing as implemented in the R package ape (Paradis et al. 2004). A correlation analysis between node depth and bootstrap values was executed with R

338 3.1.3 and graphically represented by using the packages corrplot (Wei 2013) and ggplot2 (Wickham
339 2009).

340 To quantify and measure locus dropout, we calculated the numbers of loci shared between at
341 least one individual of both sister lineages originating from each node, and divided this value by the
342 number of taxa originating from the node in question. The latter standardization was done because
343 the number of taxa varied widely between the lineages, and the probability to recover a locus
344 increases with increased hierarchical redundancy. We considered this the best measure (in
345 phylogenetic sense) of locus dropout, because loci found only in one of the sister lineages do not
346 contain phylogenetically useful information and therefore fall into the locus dropout zone. The
347 count of loci is affected also by the quality of the sample. Variation in sample quality results in
348 increased variance, which may complicate observing true patterns. To test if the data are consistent
349 with the predicted linear locus decay (Fig. 1), we fitted a linear regression model (function 'lm' in R
350 3.2.2) to the data on number of loci and the corresponding node depth values. Confidence intervals
351 were derived for the regression slope (function 'confint') and fitted regression line (function
352 'predict.lm'). Potential deviation from the linear locus decay was investigated by comparing the
353 linear regression model to a quadratic regression fitted with the same function. Linear and quadratic
354 regression models were compared on the grounds of AIC and BIC, but we also used the coefficient
355 of determination (R^2; given by the R function 'lm') in assessing model explanatory power.

356 To examine SNP and PIS dropouts, only SNPs/PIS of loci recovered in both sister lineages of
357 each node at least once were considered. To eliminate the effects of hierarchical redundancy, these
358 numbers of SNPs/PIS were divided by the number of taxa found at lineages originating from each
359 node. To test if the number of SNPs peak at intermediate node depth values (Fig. 1), we fitted a
360 quadratic regression model (R function 'lm') to the data on numbers of SNPs and corresponding
361 node depth values. Confidence intervals for the coefficient for squared node depth and the fitted
362 regression curve were derived as above. The presence of a peak in the number of SNPs along node

Commented [KLJ9]: discussion, not method (already discussed in discussion). How: To eliminate the effect of different quality between samples, we tested the data consistency… or something like this.

depth axis was further assessed by comparing the quadratic regression model to a linear one on the grounds of AIC and BIC, and by examining the R^2 values of the two models. The analysis for PIS was conducted otherwise similarly as for SNP dropout, except that the number of PIS per taxon was logarithmically transformed as ln([number of PIS) + 1]) (one added because data include zeros) to ensure model goodness-of-fit.

The effect of branch length was controlled for assessing the contribution of SNPs, PIS, and loci to node support., We first modelled the dependence of bootstrap values on branch length with an asymptotic non-linear regression through the origin (self-starting regression function 'SSasympOrig' in the R function 'nls'). Observations were weighted with the number of SNPs for the analysis of SNP and PIS contribution to node support (PIS include zeros, precluding its use as weights, but the number of PIS is strongly and positively correlated with number of SNPs; see below), and with the number of loci for the assessment of the contribution of loci to node support. The contribution of SNPs, PIS, and loci to node support was analyzed separately because the numbers of SNPs, PIS, and loci are strongly and positively correlated (Pearson's correlations [r]: $r_{SNP\text{-}PIS} = 0.957$, $t_{39} = 20.5$, $P < 0.0001$; $r_{SNP\text{-}loci} = 0.898$, $t_{39} = 12.7$, $P<0.0001$; $r_{PIS\text{-}loci} = 0.781$, $t_{39} = 7.80$, $P < 0.0001$). We took residuals from the above non-linear asymptotic regression models and used them as response variables (i.e. the component of node support not explained by branch length; hereafter called as bootstrap residuals) in subsequent analyses. Variation in the bootstrap residuals was analyzed with linear models (R function 'lm') where node depth and either the number of SNPs, the number of PIS, or number of loci were the explanatory variables. Interaction between the explanatory variables was included in both models.

RESULTS

Optimization of ddRAD loci parameters

388 On average, approximately five million reads per individual were obtained, of which 82.3% were
389 retained after stringent quality filtering steps (Table 1). After filtering and clustering, the ddRADseq
390 data matrix yielded approximately 16,000 loci per specimen, with a minimum coverage of 3x after
391 filtering for paralogs (Table 1; Table S3). Only two loci (90 and 98 nucleotides) originated from
392 *Wolbachia pipientis*.

393 The total number of loci ranged from 10 to 8,737 between the nine data matrices, demonstrating
394 the dramatic effect of parameter selection on the amount of data (Table 2). No shared loci were
395 recovered across all 43 individuals in any of the data matrices, and only one locus was retained
396 across 24 individuals (Table S4). Data assemblages that maximized the number of individuals per
397 locus contained relatively few loci and SNPs, but at the same time reduced the amount of missing
398 data. Those matrices produced discordant phylogenies compared to those with lower value of *m*
399 (e.g. ddRAD-*c*85*m*21 phylogeny; Fig.S3f). The different clustering thresholds had a significant
400 effect on the total number of loci (range 794–3,833 loci), variable sites (range 18,001–224,916) as
401 well as the PIS (range 5,122–69,029) (Table 2). The pairwise p-distance between specimens ranged
402 from 0.1% and 14.7% across all specimens and data matrices, and showed that the parameters of
403 both *m* and clustering thresholds (*c*) have a significant effect on mean distances between the
404 specimens (Fig. S4). Resulting data matrices analyzed in RAxML produced overall similar tree
405 topologies for most trials, but ddRAD-*c*85*m*21 produced a poorly resolved and very deviant tree
406 probably as a result of scarcity of retained loci (Fig S3). The tree based on the strictest clustering
407 threshold (ddRAD-*c*95*m*6) also differed considerably from the other trees. In that tree, the number
408 of SNPs was higher than in ddRAD-*c*85*m*12 and comparable to ddRAD-*c*85*m*9, but the proportion
409 of missing data was clearly higher (Fig S3).

410 *Phylogeny of Eupithecia*

411 Of ddRAD phylogenies, the one based on ddRAD-*c*85*m*6 data (726,658 bp) was selected for
412 further comparisons because of its general congruence with several other data sets and high number

of retained loci (3,256) and SNPs (3,164). Phylogenetic trees based on other data matrices of ddRAD are provided in the Supplementary Material (Fig. S3) and basic statistics in Table 2. Concatenated nuclear and mitochondrial Sanger data included 6,172 bp and 8 loci. (Table 2, Fig. 2).

The ddRAD and Sanger phylogenies were similar but not identical, the ddRAD data providing better support than Sanger data from intermediate to shallow nodes (bootstrap mostly 100% at < 0.45 depth; see Fig. 3a), whereas both ddRAD and Sanger data showed moderate to poor resolution at deeper-level nodes (at > 0.45 depth). The mt COI phylogeny produced a poorly resolved tree with low bootstrap values at most of the nodes, and the bootstrap values dropped especially fast between 0.2 to 0.4 depth (Fig. 3b, Fig. S3i).

The ddRAD topology suggests that *E. abietaria* is the sister taxon to all other sampled *Eupithecia*, while the Sanger topology places *E. actaeata* in that position, indicating a clear conflict between the data sets (Fig. 2). The positions of *E. centaureata*, *E. immundata* and *E. irriguata* remain largely unclear. *E. simpliciata* clustered with *E. semigraphata* in the ddRAD topology (bootstrap 100%; Fig. 2a), while it grouped (although poorly supported) with *E. satyrata*, *E. indigata*, *E. conterminata*, and *E. intricata* in the Sanger topology (bootstrap 36%; Fig. 2b). *E. simpliciata* and *E. semigraphata* shared 97 ddRAD loci, whereas *E. simpliciata* shared only two ddRAD loci with *E. satyrata*, *E. indigata*, *E. conterminata* and *E. intricata* (Fig. S5). *Eupithecia vulgata* also showed a conflict between ddRAD and Sanger datasets. The number of recovered loci of *E. vulgata* was 107, being the lowest of all species in the ddRAD dataset (Table 1, Fig. S6). In a trial with, *E. tantillaria* and *E. vulgata* removed, having highest levels of missing data, the phylogenetic placement and relationships of the species showing conflict between ddRAD and Sanger data (e.g., *E. semigraphata*, *E. simpliciata*) remained the same (see Fig. S7b). The exclusion of the six poorest-quality samples did not significantly affect the phylogenetic results.

Of the reference assembly, an average of 271,114 reads per sample were mapped to the 26 reference genomes of Lepidoptera, while an average of 286,552 reads per sample remained unmapped (Table

S3). After filtering, an average of 31,748 clusters per sample were obtained , with an average of 32.4 per sample for cluster depth. The final dataset from the reference assembly consisted of 822 recovered loci per sample across more than three individuals. The phylogenetic hypothesis based on the reference assembly produced in a remarkably incongruent tree with both the *de novo* ddRAD assembly tree and the Sanger tree (Fig. S8).

Effects of locus conservativeness on SNP frequency

The number of SNPs per locus showed considerable variation at each value of individuals per locus (range 6-24), demonstrating pronounced variation in locus conservativeness regardless of its likelihood to be recovered. The average number of SNPs/locus, however, tended to decrease with increasing number of individuals/locus across loci shared by a minimum of 10 individuals (Fig. 4), demonstrating the connection between the locus dropout and the type of retained loci. The quadratic model (Table S5) explained the data much better than the linear model (ΔAIC=18.3, ΔBIC=12.3 in favor of the quadratic model). The 95% adjusted bootstrap percentile confidence intervals encompassed the fitted regression curve derived from the generalized linear model, lending support to inferences based on the regression model even though the normality assumption of the residuals was violated in the regression model. The number of recovered loci decreased dramatically when an increasing number of individuals were required to share a locus (Fig. S9).

Patterns of locus, SNP and PIS dropouts and their effects on node confidence

Locus dropout towards deeper nodes was linear, as expected (Table 3; Fig. 5a), the 95% confidence interval of the regression slope (-315, -46.7) and the support for the linear regression over the quadratic one (ΔAIC=1.98, ΔBIC=3.70 in favor of the linear model) supporting the prediction presented in Figure 1. The coefficients of determination were the same for both the linear (R^2 = 0.16) and quadratic (R^2 = 0.16) regression models for locus dropout, further supporting the choice of the simpler linear regression model. The number of SNPs was highest at intermediate

462 node depth and decreased towards shallow and deep nodes (Table 3; Fig. 5b), which is also
463 consistent with the prediction (cf. Fig. 1). Consistency with the prediction is further supported by
464 the 95% confidence interval of the coefficient for squared node depth (-14697, -1781), the support
465 for the quadratic regression over the linear regression model (ΔAIC=4.63, ΔBIC=2.92 in favor of
466 the quadratic model), and the higher coefficient of determination for the quadratic (R^2 = 0.30) than
467 the linear (R^2 = 0.17) regression model. The ln-transformed number of PIS linearly increased
468 towards deep nodes (Fig. 5c; 95% confidence interval of the slope: 5.29, 13.0), and the linear model
469 was supported over the quadratic one (ΔAIC=1.87, ΔBIC=3.20 in favor of the linear model), the
470 coefficients of determination being similar for both the linear (R^2 = 0.48) and quadratic (R^2 = 0.48)
471 models.

472 The effect of branch length on bootstrap values was removed by analysing variation in residuals
473 from a non-linear asymptotic regression of bootstrap values on branch length (bootstrap residuals).
474 Variation in bootstrap residuals was only explained by node depth, and not by the number of loci,
475 SNPs or parsimony informative SNPs (PIS) in ddRAD data (Table S6; Fig. 6).

Commented [KLJ10]: Method

476

477 DISCUSSION

478

479 Previous studies have demonstrated that RAD methods are generally efficient in inferring
480 shallow-level phylogenies (e.g. Tiffin and Ross-Ibarra 2014; Hou et al. 2015; Leaché et al. 2015b;
481 Ree and Hipp 2015; Andrews et al. 2016; Kim et al. 2016). RAD phylogenies have often yielded
482 unexpectedly well-resolved relationships also at deep phylogenetic levels, and even tens of millions
483 of years old divergences have been resolvable (Rubin et al. 2012; Cariou et al. 2013; Leaché et al.
484 2015a; Herrera and Shank 2016). Eaton et al. (2017) recently recognized that growing hierarchical
485 redundancy towards the deeper splits constitutes a major reason for the high power of RAD
486 methods at relatively deep phylogenetic levels. As far as we know, our study is the first to

investigate how locus dropout affects the amount of phylogenetic information at different phylogenetic depths. We demonstrate that the number of retained loci is not an accurate measure of phylogenetic information content in RAD data sets and that they tend to become more information-rich towards the deeper phylogenetic levels. Our comparison with an eight-gene Sanger data indicates that ddRAD sequencing yields overall congruent tree topologies despite a lack of retained loci that are shared among all studied taxa. While we base our conclusions on an empirical data set of 35 species of moths, the observed patterns are unlikely to be special to this particular moth group, but are likely to occur in the RAD data sets in other taxa as well.

Commented [KLJ11]: I would delete these words

Effects of sample quality and the adopted protocol

Relatively low number (mean 610) of consensus loci was retained in the ddRAD data set with minimum number of individuals per locus value of 6. While an age estimate for the genus is not available, it is likely that it is less than 10-20 million years old, given that a deep split within the subfamily to which *Eupithecia* belongs to is estimated at 33 million years ago (Wahlberg et al. 2013). We observed a very strong locus dropout effect as demonstrated by the observation that while on average 16k loci were recovered per specimen, none of them was recovered across all specimens.

Do *Eupithecia* represent a phylogenetic borderline-case for RAD methods being efficient? The power of the analysis could likely be substantially increased by improving sample quality, repeating the ddRAD library preparation, using different (or additional) restriction enzymes, using a different RAD method, and increasing sampling intensity. Optimally, samples to be used should be stored in a way that minimizes the degradation of DNA as the level of DNA degradation is directly correlated with the probability of finding a given locus. For practical reasons, like in our case, samples of suboptimal quality may be included as the availability of alcohol or freezer-preserved samples is usually limited, to increase density of taxon sampling. In some cases, the final number of retained

Commented [KLJ12]: How does this sentence link to the current shape of the text? This is only answered in abstract now.

Commented [KLJ13]: above in some references it is stated that degradation is not necessarily a problem as fragments useful need not be long. So, is there incongruence worth mentioning?

512 loci remained much lower than in others. This could have been partly avoided by increasing the
513 amount of tissue used for DNA extraction, but for very small species (the majority of extant species
514 are small) even this is not an option. A substantial increase in the amount of loci could have been
515 obtained by duplication of the RAD library preparation. This is supported by the observation that,
516 on average, only 20.6% of all loci showed a depth value of at least 3 and could be retained (Table
517 S7). Furthermore, since a majority of loci were recovered less than four times, many loci not falling
518 within the locus dropout zone due to mutation-disruption were likely not recovered even a single
519 time. The power of RAD analysis could additionally be increased by repeating the analysis with
520 another set of restriction enzymes, although this nearly duplicates the costs, a reason for which such
521 trials are rare. Additionally, single digest RAD methods yield more phylogenetic information than
522 double-digest methods such as the one used here (Andrews et al. 2016). Finally, the tree resolution
523 could be improved by a denser and more balanced taxon sampling (Eaton et al. 2017), and
524 especially by the inclusion of "critical" taxa, namely those cutting the long branches of the tree and
525 hence increasing the hierarchical redundancy of the data.

526 Due to the low DNA quantity of the original DNA extracts, we conducted a whole-genome
527 amplification (WGA) for each sample. WGA may amplify different parts of the genome in a biased
528 way and introduce errors in the amplified regions (Pinard et al. 2006; Blair et al. 2015; Burford
529 Reiskind et al. 2016). On the contrary, WGA produced accurate reduced representations of human,
530 mouse and bird genomes (Barker et al. 2004; Han et al. 2012; Rheindt et al. 2014). Tin et al. (2014)
531 conducted WGA for RAD tags with ant museum material with degraded DNA, and similarly
532 observed no significant genomic bias due to the genomic enrichment. If WGA under-amplifies the
533 genome, a lower number of unique loci and a greater coverage of the amplified regions is expected.
534 Alternatively, if WGA introduces errors to amplified regions, an exaggerated degree of SNPs is
535 expected. We attempted to validate our data through careful bioinformatics scrutiny and applied a

strict *m* (minimum number of individuals per locus) value, albeit at the expense of a number of loci included in the final data set.

Effects of clustering threshold and minimum individual parameters on RAD data matrix

Although on average approximately 16,000 loci for each sample were recovered for *Eupithecia*, an average of only 610 loci per individual were retained in the final data set. This represents a well-demonstrated drawback of RAD methods. For example, Rheindt et al. (2014) could save only 2.9-3.9% of all recovered SNPs in their between-population analyses. The breadth of the RAD data is greatly affected by the stringency of clustering and minimum individual thresholds. Negligence in these steps may easily lead to the inclusion of paralogs, contaminant reads and otherwise misleading data, reducing the overall reliability of data. RAD methods have a benefit of being feasible for non-model taxa lacking a reference genome, but the reverse side of this is that filtering out alien reads and paralogs is complicated and must be done informatically (Ree and Hipp 2015).

We assessed the effects of both the clustering threshold and the minimum individual threshold on the tree topology of each data matrix. Most of our analyses based on ddRADseq matrices produced congruent trees with high support values for most nodes. The minimum individual parameter in particular controls the amount of missing data as it has a direct relation with the number of loci (or SNPs) in the final matrix (Ree and Hipp 2015). The variation in the degree of missing data did not strongly affect the tree topologies, but the largest, and thus most informative, data matrices resulted in the highest phylogenetic support for nodes (see Table 2; Fig. S3). This result is consistent with previous observations that large amounts of missing data in RADseq data sets do not adversely affect the accuracy of phylogenetic inference (Rubin et al. 2012; Keller et al. 2013; Hipp et al. 2014; Takahashi et al. 2014; Hou et al. 2015; Herrera and Shank 2016). However, Leaché et al. (2015a) demonstrated that, although this generally holds true, data sets with high levels of missing data are error-prone. They emphasized that the statistical node support value is not

equal to its true confidence (see also Rubin et al. 2012), but may artificially result from biases of the data. In our case, broad congruence between the two phylogenies based on independent data sets suggest that missing data did not have significant adverse effects on recovering the true tree topology.

Comparison of RAD and Sanger tree topologies

Previous comparisons between Sanger and RAD data sets have shown that RAD data generally outperform Sanger data sets (Eaton and Ree 2013; Keller et al. 2013; Cruaud et al. 2014; Escudero et al. 2014; Hipp et al. 2014; Herrera et al. 2015; Ruane et al. 2015). In our case, the ddRAD and Sanger data provided overall similar tree topologies. This would be an unlikely output if one or both of the data sets were poor of phylogenetic information and hence misleading. However, a few remarkable cases of incongruence were detected. In both trees, some of the deeper nodes were statistically poorly supported likely due to very short internodal branches. Nodes at intermediate phylogenetic depth were better supported by ddRAD data compared to Sanger data, but at the deepest levels bootstrap values in ddRAD data sets dropped steeply (Fig. 3). A likely explanation for this is the decay of phylogenetic information due to the dropout of data (Fig. 5).

Based on ddRAD data, the sister species to the rest of the sampled *Eupithecia* is *E. abietaria*. Although no prior rigorous analysis of phylogenetic relationships in *Eupithecia* exists to support this finding, we find it a likely scenario based on the morphological distinctiveness of this taxon within *Eupithecia* but shared with *Pasiphila*, our outgroup taxon. Using Sanger data, the species in this basal position was inferred to be *E. actaeata*, a species that shows close overall morphological similarity with many other species of *Eupithecia*. However, in Sanger data the monophyly of the sampled *Eupithecia* with *E. actaeata* excluded is very strongly supported, whereas in ddRAD data the monophyly of all except for *E. abietaria* remains supported by a bootstrap support (BS) of only 68%. This incongruence is difficult to explain, since *E. actaeata* is firmly (100% BS) associated

with two other species (*E. exiguata* and *E. assimilata*) in all ddRAD trials and is never placed even close to the root.

Another remarkable case of incongruence between the data sets is the position of *E. simpliciata*, which appears as a highly unstable taxon whose position is poorly supported in the Sanger data, and separated by a very short internodal branch. In the ddRAD data, it associates with *E. semigraphata* with 100% BS, and together with three other species (*E. millefoliata*, *E. icterata* and *E. denotata*), forms a strongly supported entity, which, with the exclusion of *E. simpliciata*, is also strongly supported by Sanger data as well. Interestingly, all these five species share an ecological trait, their flying period being late summer. We conclude pattern displayed by *E. simpliciata* in Sanger data to be likely caused by a shortage of phylogenetic information in this data set, which, unlike ddRAD data, performs poorly at intermediate phylogenetic levels (Fig. 3).

The position of *E. vulgata* represents another remarkable case of incongruence between the data sets. On the basis of morphology, this species appears to be a close relative of *E. assimilata*, with which it associates in Sanger data with strong support (together with *E. exiguata*). In contrast, *E. vulgata* associates with *E. selinata* in the ddRAD tree. The position of *E. vulgata* is, however, significantly unstable in the various ddRAD trials (Fig. S3). The reason lies in the poor success of *E. vulgata* for loci recovery. With a low number of loci recovered (107) and a mean locus coverage of as high as 854, *E. vulgata* represents a likely case of poor quality in the original DNA template..

Patterns of loci, SNPs and PIS in RAD datasets

Huang and Knowles (2016) demonstrated with simulations that the proportion of missing data is associated with the type of loci retained in the data. This is intuitively plausible as it can be expected that slowly evolving loci are less likely to drop out than rapidly evolving loci. Our study is the first to demonstrate with empirical data that the more often a locus is found among species, the

poorer they are in phylogenetic information (measured in this analysis by SNPs). Likely for the same reason, the minimum number of individuals per locus value (*m*) is negatively correlated with the pairwise genetic distance between specimens. While the negative correlation between the locus recovery rate and their SNP content was statistically highly significant, there is overall much variation in SNP frequency, and the observed decline of SNPs is not steep. We presume that this effect is mitigated by opposite effects: conservative loci are more "long-living" (less sensitive to mutation-disruption), thus have had a longer time to accumulate mutations. These opposite effects might even compensate each other. The observed trend may therefore actually be explained by the higher proportion of ultra-conserved loci retained with higher values of individuals/locus. Figure 4 suggests that this might be the case, since at high values of individuals/locus very conservative loci are present, while loci with over 80 SNPs are not found with >15 individuals/locus.

Locus dropout is caused by the disruption of restriction as a result of mutation at the restriction region, resulting in a pattern of decline in locus sharing with phylogenetic distance. Accordingly, in our data, the number of loci shows a constant decline along with increased coalescence time (node depth), and nearly reaches zero at the deepest nodes. As we hypothesized, the number of loci does represents a good proxy for phylogenetic information (number of SNPs and PIS) retained in the data (Figs. 5b and 5c). The shallow nodes with large numbers of shared loci between the sister lineages were constantly poor of SNPs and PIS in relation to the sister lineages at the intermediate phylogenetic levels, highlighting that the number of loci is not supposed to be a good proxy of information content of the data in the population genetic studies. The number of SNPs is also low in the deepest phylogenetic nodes. This results directly from the decay of recovered loci. While the loci retained at the deepest levels tend to be conservative, they are not necessarily particularly poor in phylogenetic information because they have had the longest time to accumulate mutations, as suggested by the relatively high number of PIS in the deepest phylogenetic nodes.

Interestingly, neither the number of loci or SNPs, nor PIS explained node support when the confounding effect of the length of the branch leading to the node was eliminated. Only node depth explained node support. The lack of contribution to node support should, however, be considered with caution, because our data do not contain much information about these effects. Our observations are strongly biased towards low numbers of loci, SNPs and PIS (see Fig. 6). Secondly, the observed bootstrap supports are strongly dominated by very high values, which also makes it difficult to estimate the dependency of node support on any explanatory variables. Furthermore, bootstrap values do not provide an accurate estimate of the true phylogeny under all conditions (refs). Owing to these reasons, we cannot exclude the possibility that the number of loci, and the number of SNPs or PIS in particular, are positively correlated with the node confidence, as would be expected. Yet, given the clear-cut results concerning locus and SNP/PIS dropouts, any data are predicted to be unevenly spread in the node depth-phylogenetic information (numbers of loci/SNPs/PIS) space, which remains a potential challenge for future analyses.

Commented [MM14]: https://academic.oup.com/sysbio/article-abstract/42/2/182/1730933

CONCLUSIONS

RAD methods are characterized by large numbers of recovered loci combined with a strong locus dropout effect and large proportions of missing data, arguably compromising their use at deep phylogenetic levels. The plain number of retained loci, however, does not provide a good proxy for the amount of phylogenetic information in the data, because (i) retained loci tend to become more informative towards the deeper phylogenetic levels (Huang and Knowles 2016, this study), (ii) hierarchical redundancy is increased towards deeper phylogenetic levels (Eaton et al. 2017), and (iii) the number of loci does not equal the number of SNPs and PIS (this study). Thus, attention should be paid to available phylogeny-informative SNPs retained at different phylogenetic depths. Comprehensive and balanced taxon sampling helps resolving phylogenetic affinities also at relatively deep phylogenetic levels. We demonstrated this with a comparison of ddRAD and

multigene Sanger-sequencing based phylogeny in 35 species of a diverse moth genus. The number of available loci could substantially be further increased by repeating the library preparation and applying different restriction enzymes.

Acknowledgments

We are grateful to Laura Törmälä and Soile Alatalo for their efficient work in lab and for continuously developing laboratory protocols and practices. We are grateful to the two anonymous reviewers and Vlad Dinca for providing numerous useful comments on the manuscript. The authors also wish to acknowledge CSC – IT Center for Science, Finland for providing computational resources. This study was financially supported by the Academy of Finland through a research grant #277984 to MM. NW acknowledges support from the Swedish Research Council and SMK thanks Emil Aaltonen Foundation and the Estonian Research Council (grant PUT1474) for research grants. We also would like to thank people at the Canadian Centre for DNA Barcoding (CCDB) for sequencing the DNA barcode regions and continuous support with BOLD data management.

Supplementary Material

Data are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.474nd.

Field Code Changed

References

Andrews K.R., Good J.M., Miller M.R., Luikart G., Hohenlohe P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17:81–92.

Arnold B., Corbett-Detig R.B., Hartl D., Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol. Ecol. 22:3179–3190.

Baird N., Etter P., Atwood T., Currey M., Shiver A., Lewis Z., Selker E., Cresko W., Johnson E.

683 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One.
684 3:e3376.

685 Barker D.L., Hansen M.S.T., Faruqi A.F., Giannola D., Irsula O.R., Lasken R.S., Latterich M.,
686 Makarov V., Oliphant A., Pinter J.H., Shen R., Sleptsova I., Ziehler W., Lai E. 2004. Two
687 methods of whole-genome amplification enable accurate genotyping across a 2320-SNP
688 linkage panel. Genome Res. 14:901–7.

689 Blair C., Campbell C.R., Yoder A.D. 2015. Assessing the utility of whole genome amplified DNA
690 for next-generation molecular ecology. Mol. Ecol. Resour. 15:1079–1090.

691 Brandley M.C., Bragg J.G., Singhal S., Chapple D.G., Jennings C.K., Lemmon A.R., Moriarty
692 Lemmon E., Thompson M.B., Moritz C. 2015. Evaluating the performance of anchored hybrid
693 enrichment at the tips of the tree of life: a phylogenetic analysis of Australian Eugongylus
694 group scincid lizards. BMC Evol. Biol. 15:62.

695 Breinholt J.W., Earl C., Lemmon A.R., Moriarty Lemmon E., Xiao L., Kawahara A.Y., Lemmon
696 E.M., Xiao L., Kawahara A.Y. 2017. Resolving relationships among the megadiverse
697 butterflies and moths with a novel pipeline for anchored phylogenomics. Syst. Biol. 7:syx048.

698 Brunet B.M.T., Blackburn G.S., Muirhead K., Lumley L.M., Boyle B., Lévesque R.C., Cusson M.,
699 Sperling F.A.H. 2017. Two's company, three's a crowd: new insights on spruce budworm
700 species boundaries using genotyping-by-sequencing in an integrative species assessment
701 (Lepidoptera: Tortricidae). Syst. Entomol. 42:317–328.

702 Burford Reiskind M.O., Coyle K., Daniels H. V., Labadie P., Reiskind M.H., Roberts N.B., Roberts
703 R.B., Schaff J., Vargo E.L. 2016. Development of a universal double-digest RAD sequencing
704 approach for a group of nonmodel, ecologically and economically important insect and fish
705 taxa. Mol. Ecol. Resour. 16:1303–1314.

706 Canty A., Ripley B. 2015. boot: Bootstrap R (S-plus) functions. R package version 1.3-17.

Cariou M., Duret L., Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecol. Evol. 3:846–852.

Cruaud A., Gautier M., Galan M., Foucaud J. 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. Mol. Biol. Evol. 31:1272–1274.

DaCosta J.M., Sorenson M.D. 2016. ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories. Mol. Phylogenet. Evol. 94:122–135.

Dasmahapatra K.K., Walters J.R., Briscoe A.D., Davey J.W., Whibley A., Nadeau N.J., Zimin A. V., Hughes D.S.T., Ferguson L.C., Martin S.H., Salazar C., Lewis J.J., Adler S., Ahn S.-J., Baker D.A., Baxter S.W., Chamberlain N.L., Chauhan R., Counterman B.A., Dalmay T., Gilbert L.E., Gordon K., Heckel D.G., Hines H.M., Hoff K.J., Holland P.W.H., Jacquin-Joly E., Jiggins F.M., Jones R.T., Kapan D.D., Kersey P., Lamas G., Lawson D., Mapleson D., Maroja L.S., Martin A., Moxon S., Palmer W.J., Papa R., Papanicolaou A., Pauchet Y., Ray D.A., Rosser N., Salzberg S.L., Supple M.A., Surridge A., Tenger-Trolander A., Vogel H., Wilkinson P.A., Wilson D., Yorke J.A., Yuan F., Balmuth A.L., Eland C., Gharbi K., Thomson M., Gibbs R.A., Han Y., Jayaseelan J.C., Kovar C., Mathew T., Muzny D.M., Ongeri F., Pu L.-L., Qu J., Thornton R.L., Worley K.C., Wu Y.-Q., Linares M., Blaxter M.L., Ffrench-Constant R.H., Joron M., Kronforst M.R., Mullen S.P., Reed R.D., Scherer S.E., Richards S., Mallet J., Owen McMillan W., Jiggins C.D. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 487:94–98.

Davey J.L., Blaxter M.W. 2010. RADseq: Next-generation population genetics. Brief. Funct. Genomics. 9:416–423.

DeWaard J.R., Ivanova N.V., Hajibabaei M., Hebert P.D.N. 2008. Assembling DNA barcodes: analytical protocols. In: Martin C., editor. Methods in molecular biology: environmental

731 genetics. Totowaa, NJ: Humana Press. p. 275–294.

732 Eaton D.A.R. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses.
733 Bioinformatics. 30:1844–1849.

734 Eaton D.A.R., Overcast I. 2016. ipyrad: interactive assembly and analysis of RADseqdata sets.
735 Available from http://ipyrad.readthedocs.io/.

736 Eaton D.A.R., Ree R. 2013. Inferring phylogeny and introgression using RADseq data: an example
737 from flowering plants (Pedicularis: Orobanchaceae). Syst. Biol. 62:689–706.

738 Eaton D.A.R., Spriggs E.L., Park B., Donoghue M.J. 2017. Misconceptions on missing data in
739 RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst. Biol. 66:399–
740 412.

741 Ebel E.R., DaCosta J.M., Sorenson M.D., Hill R.I., Briscoe A.D., Willmott K.R., Mullen S.P. 2015.
742 Rapid diversification associated with ecological specialization in Neotropical Adelpha
743 butterflies. Mol. Ecol. 24:2392–2405.

744 Edgar R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput.
745 Nucleic Acids Res. 32:1792–1797.

746 Emerson K.J., Merz C.R., Catchen J.M., Hohenlohe P.A., Cresko W.A., Bradshaw W.E., Holzapfel
747 C.M. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc.
748 Natl. Acad. Sci. U. S. A. 107:16196–16200.

749 Escudero M., Eaton D.A.R., Hahn M., Hipp A.L. 2014. Genotyping-by-sequencing as a tool to infer
750 phylogeny and ancestral hybridization: A case study in Carex (Cyperaceae). Mol. Phylogenet.
751 Evol. 79:359–367.

752 Gonen S., Bishop S.C., Houston R.D. 2015. Exploring the utility of cross-laboratory RAD-
753 sequencing datasets for phylogenetic analysis. BMC Res. Notes. 8:299.

754 Gratton P., Trucchi E., Trasatti A., Riccarducci G., Marta S., Allegrucci G., Cesaroni D., Sbordoni

755 V. 2016. Testing classical species properties with contemporary data: How “bad species” in
756 the brassy ringlets (Erebia tyndarus complex, Lepidoptera) turned good. Syst. Biol. 65:292–
757 303.

758 Hall T. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program
759 for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41:95–98.

760 Hamilton C.A., Lemmon A.R., Moriarty Lemmon E., Bond J.E. 2016. Expanding anchored hybrid
761 enrichment to resolve both deep and shallow relationships within the spider tree of life. BMC
762 Evol. Biol. 16:212.

763 Han T., Chang C., Kwekel J. 2012. Characterization of whole genome amplified (WGA) DNA for
764 use in genotyping assay development. BMC Genomics. 13:217.

765 Heikkilä M., Mutanen M., Wahlberg N., Sihvonen P., Kaila L. 2015. Elusive ditrysian phylogeny:
766 an account of combining systematized morphology with molecular data (Lepidoptera). BMC
767 Evol. Biol. 15:260.

768 Herrera S., Shank T.M. 2016. RAD sequencing enables unprecedented phylogenetic resolution and
769 objective species delimitation in recalcitrant divergent taxa. Mol. Phylogenet. Evol. 100:70–
770 79.

771 Herrera S., Watanabe H., Shank T.M. 2015. Evolutionary and biogeographical patterns of barnacles
772 from deep-sea hydrothermal vents. Mol. Ecol. 24:673–689.

773 Hipp A.L., Eaton D.A.R., Cavender-Bares J., Fitzek E., Nipper R., Manos P.S. 2014. A framework
774 phylogeny of the American oak clade based on sequenced RAD data. PLoS One. 9:e93975.

775 Hohenlohe P.A., Amish S.J., Catchen J.M., Allendorf F.W., Luikart G. 2011. Next-generation RAD
776 sequencing identifies thousands of SNPs for assessing hybridization between rainbow and
777 westslope cutthroat trout. Mol. Ecol. Resour. 11:117–122.

778 Hou Y., Nowak M.D., Mirré V., Bjorå C.S., Brochmann C., Popp M. 2015. Thousands of RAD-seq

779 loci fully resolve the phylogeny of the highly disjunct arctic-alpine genus Diapensia
780 (Diapensiaceae). PLoS One. 10:e0140175.

781 Huang H., Knowles L.L. 2016. Unforeseen consequences of excluding missing data from next-
782 generation sequences: simulation study of RAD sequences. Syst. Biol. 65:357–365.

783 Jones J.C., Fan S., Franchini P., Schartl M., Meyer A. 2013. The evolutionary history of
784 Xiphophorus fish and their sexually selected sword: A genome-wide approach using restriction
785 site-associated DNA sequencing. Mol. Ecol. 22:2986–3001.

786 Kai W., Nomura K., Fujiwara A., Nakamura Y., Yasuike M., Ojima N., Masaoka T., Ozaki A.,
787 Kazeto Y., Gen K., Nagao J., Tanaka H., Kobayashi T., Ototake M. 2014. A ddRAD-based
788 genetic map and its integration with the genome assembly of Japanese eel (Anguilla japonica)
789 provides insights into genome evolution after the teleost-specific genome duplication. BMC
790 Genomics. 15:233.

791 Keller I., Wagner C.E., Greuter L., Mwaiko S., Selz O.M., Sivasundar A., Wittwer S., Seehausen O.
792 2013. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation
793 in the rapid radiation of Lake Victoria cichlid fishes. Mol. Ecol. 22:2848–2863.

794 Kim C., Guo H., Kong W., Chandnani R., Shuang L.-S., Paterson A.H. 2016. Application of
795 genotyping by sequencing technology to a variety of crop breeding programs. Plant Sci.
796 242:14–22.

797 Kozlov M. V., Mutanen M., Lee K.M., Huemer P. 2017. Cryptic diversity in the long-horn moth
798 Nemophora degeerella (Lepidoptera: Adelidae) revealed by morphology, DNA barcodes and
799 genome-wide ddRAD-seq data. Syst. Entomol. 42:329–346.

800 Leaché A.D., Banbury B.L., Felsenstein J., de Oca A. nieto-M., Stamatakis A. 2015a. Short tree,
801 long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP
802 phylogenies. Syst. Biol. 64:1032–1047.

803 Leaché A.D., Chavez A.S., Jones L.N., Grummer J.A., Gottscho A.D., Linkem C.W. 2015b.
804 Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus
805 restriction site associated DNA sequencing. Genome Biol. Evol. 7:706–719.

806 Leaché A.D., Fujita M.K., Minin V.N., Bouckaert R.R. 2014. Species delimitation using genome-
807 wide SNP Data. Syst. Biol. 63:534–542.

808 Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored Hybrid Enrichment for Massively
809 High-Throughput Phylogenomics. Syst. Biol. 61:727–744.

810 Lemmon E.M., Lemmon A.R. 2013. High-throughput genomic data in systematics and
811 phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44:99–121.

812 Mardis E.R. 2013. Next-generation sequencing platforms. Annu. Rev. Anal. Chem. 6:287–303.

813 McCluskey B., Postlethwait J. 2015. Phylogeny of zebrafish, a "model species," within Danio, a
814 "model genus." Mol. Biol. Evol. 32:635–652.

815 McDunnough J.H. 1949. Revision of the North American species of the genus Eupithecia
816 (Lepidoptera, Geometridae). Bull. Am. Museum Nat. Hist. 93:533–734.

817 Miller M.R., Dunham J.P., Amores A., Cresko W.A., Johnson E.A. 2007. Rapid and cost-effective
818 polymorphism identification and genotyping using restriction site associated DNA (RAD)
819 markers. Genome Res. 17:240–248.

820 Mironov V. 2003. Larentiinae II: Perizomini, Eupitheciini. In: Hausmann A, ed. The Geometrid
821 Moths of Europe 1, Apollo Books, Stenstrup, 463 pp. .

822 Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri
823 T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer
824 A.J., Aspock U., Aspock H., Bartel D., Blanke A., Berger S., Bohm A., Buckley T.R., Calcott
825 B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin
826 L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu

827 H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-
828 Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schutte
829 K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan
830 M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune
831 T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K.F., Wu Q., Wu G., Xie
832 Y., Yang S., Yang Q., Yeates D.K., Yoshizawa K., Zhang Q., Zhang R., Zhang W., Zhang Y.,
833 Zhao J., Zhou C., Zhou L., Ziesmann T., Zou S., Li Y., Xu X., Zhang Y., Yang H., Wang J.,
834 Wang J., Kjer K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect
835 evolution. Science (80-.). 346:763–767.

836 Mutanen M., Wahlberg N., Kaila L. 2010. Comprehensive gene and taxon coverage elucidates
837 radiation patterns in moths and butterflies. Proc. Biol. Sci. 277:2839–2848.

838 Nadeau N.J., Martin S.H., Kozak K.M., Salazar C., Dasmahapatra K.K., Davey J.W., Baxter S.W.,
839 Blaxter M.L., Mallet J., Jiggins C.D. 2013. Genome-wide patterns of divergence and gene flow
840 across a butterfly radiation. Mol. Ecol. 22:814–826.

841 Pante E., Abdelkrim J., Viricel A., Gey D., France S.C., Boisselier M.C., Samadi S. 2015. Use of
842 RAD sequencing for delimiting species. Heredity (Edinb). 114:450–459.

843 Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R
844 language. Bioinformatics. 20:289–290.

845 Peterson B.K., Weber J.N., Kay E.H., Fisher H.S., Hoekstra H.E. 2012. Double digest RADseq: an
846 inexpensive method for de novo SNP discovery and genotyping in model and non-model
847 species. PLoS One. 7:e37135.

848 Pinard R., De Winter A., Sarkis G.J., Gerstein M.B., Tartaro K.R., Plant R.N., Egholm M.,
849 Rothberg J.M., Leamon J.H. 2006. Assessment of whole genome amplification-induced bias
850 through high-throughput, massively parallel whole genome sequencing. BMC Genomics.

7:216.

Puritz J.B., Matz M. V, Toonen R.J., Weber J.N., Bolnick D.I., Bird C.E. 2014. Demystifying the RAD fad. Mol. Ecol. 23:5937–5942.

R Core Team. 2015. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.r-project.org/.

Ree R., Hipp A. 2015. Inferring phylogenetic history from restriction site associated DNA (RADseq). In: Hörandl E., Appelhans M., editors. Next Generation Sequencing in Plant Systematics. Koenigstein: Koeltz Scientific Books. p. 181–204.

Regier J.C., Mitter C., Zwick A., Bazinet A.L., Cummings M.P., Kawahara A.Y., Sohn J.-C., Zwickl D.J., Cho S., Davis D.R., Baixeras J., Brown J., Parr C., Weller S., Lees D.C., Mitter K.T. 2013. A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). PLoS One. 8:e58568.

Rheindt F.E., Fujita M.K., Wilton P.R., Edwards S.V. 2014. Introgression and phenotypic assimilation in zimmerius flycatchers (Tyrannidae): Population genetic and phylogenetic inferences from genome-wide SNPs. Syst. Biol. 63:134–152.

Rowe H.C., Renaut S., Guggisberg A. 2011. RAD in the realm of next-generation sequencing technologies. Mol. Ecol. 20:3499–3502.

Ruane S., Raxworthy C., Lemmon A. 2015. Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: an empirical study on Malagasy. BMC Evol. Biol. 15:221.

Rubin B.E.R., Ree R.H., Moreau C.S. 2012. Inferring phylogenies from RAD sequence data. PLoS One. 7:e33394.

Scoble M.J., Hausmann A. 2007. Online list of valid and available names of the Geometridae of the World. Available from http://www.lepbarcoding.org/geometridae/species_checklists.php.

875 Sihvonen P., Mutanen M., Kaila L., Brehm G., Hausmann A., Staude H.S. 2011. Comprehensive
876 molecular sampling yields a robust phylogeny for geometrid moths (Lepidoptera:
877 Geometridae). PLoS One. 6:e20356.

878 Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
879 thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.

880 Streicher J.W., Devitt T.J., Goldberg C.S., Malone J.H., Blackmon H., Fujita M.K. 2014.
881 Diversification and asymmetrical gene flow across time and space: Lineage sorting and
882 hybridization in polytypic barking frogs. Mol. Ecol. 23:3273–3291.

883 Suchan T., Pitteloud C., Gerasimova N.S., Kostikova A., Schmid S., Arrigo N., Pajkovic M.,
884 Ronikier M., Alvarez N. 2016. Hybridization capture using RAD probes (hyRAD), a new tool
885 for performing genomic analyses on collection specimens. PLoS One. 11:e0151651.

886 Takahashi T., Nagata N., Sota T. 2014. Application of RAD-based phylogenetics to complex
887 relationships among variously related taxa in a species flock. Mol. Phylogenet. Evol. 80:137–
888 144.

889 Tiffin P., Ross-Ibarra J. 2014. Advances and limits of using population genetics to understand local
890 adaptation. Trends Ecol. Evol. 29:673–680.

891 Tin M.M.Y., Economo E.P., Mikheyev A.S. 2014. Sequencing degraded DNA from non-
892 destructively sampled museum specimens for RAD-tagging and low-coverage shotgun
893 phylogenetics. PLoS One. 9:e96793.

894 Venables W., Ripley B. 2002. Modern applied statistics with S. 4th edition. New York: Springer.

895 Viricel A., Pante E., Dabin W., Simon-Bouhet B. 2014. Applicability of RAD-tag genotyping for
896 interfamilial comparisons: empirical data from two cetaceans. Mol. Ecol. Resour. 14:597–605.

897 Wagner C.E., Keller I., Wittwer S., Selz O.M., Mwaiko S., Greuter L., Sivasundar A., Seehausen O.
898 2013. Genome-wide RAD sequence data provide unprecedented resolution of species

boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22:787–798.

Wahlberg N., Wheat C.W. 2008. Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic DNA extractions of Lepidoptera. Syst. Biol. 57:231–242.

Wahlberg N., Wheat C.W., Peña C. 2013. Timing and patterns in the taxonomic diversification of Lepidoptera (butterflies and moths). PLoS One. 8:e80875.

Wang X.Q., Zhao L., Eaton D. a R., Li D.Z., Guo Z.H. 2013. Identification of SNP markers for inferring phylogeny in temperate bamboos (Poaceae: Bambusoideae) using RAD sequencing. Mol. Ecol. Resour. 13:938–945.

Wei T.Y. 2013. corrplot: Visualization of a correlation matrix. Available from http://cran.r-project.org/package=corrplot.

Wickham H. 2009. ggplot2: elegant graphics for data analysis. New York: Springer.

Wilkinson M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. Mol. Biol. Evol. 13:437–444.

Zahiri R., Holloway J.D., Kitching I.J., Lafontaine J.D., Mutanen M., Wahlberg N. 2012. Molecular phylogenetics of Erebidae (Lepidoptera, Noctuoidea). Syst. Entomol. 37:102–124.

FIGURE 1. Schematic representation of actual numbers of shared loci, SNPs and PIS, and those expected to be observed in RAD data sets between two lineages along their coalescence time (starting from a coalescence time of zero). The actual number of homologous loci is constantly but slowly decreasing with increasing coalescence time. The actual number of SNPs and PIS is increasing first fast because most mutations represent new SNPs and PIS, but then at a steadily decreasing pace because of saturation of mutations at any given site. The number of loci observed in RAD data is expected to decrease at constant rate as a result of mutations accumulating to the restriction sites, finally reaching zero. This effect is called locus dropout or locus decay. The number of observed SNPs and PIS in the data are affected by their actual number and recovered number of loci, resulting in a peaked curve with an optimum at intermediate phylogenetic levels.

FIGURE 2. Phylogenetic trees of *Eupithecia* based on (a) ddRAD-*c*85*m*6 and (b) combined nuclear and mitochondrial Sanger data. The combined nuclear and mitochondrial tree was constructed based on the nuclear CAD, EF1α, GAPDH, IDH, MDH, RpS5, wingless and mitochondrial COI genes. Phylogenetic trees were inferred with RAxML with 500 bootstrap replicates. Bootstrap values are indicated near branches.

FIGURE 3. Bootstrap values in relation to node depth in (a) ddRAD-*c*80, ddRAD-*c*85, ddRAD-*c*90 and (b) combined NR+MT, mt COI. Shaded regions represent 95% confidence intervals around average coherence.

FIGURE 4. Number of SNPs per locus in relation to the number of individuals per locus. Open circles indicate the observations, and the thick and thin lines depict the fitted regression (a quadratic generalized linear model with negative binomial error distribution and a logarithmic link function) and its 95% confidence intervals, respectively. The red crosses indicate the mean numbers of SNPs per locus in each category, and the red whiskers depict the 95% adjusted bootstrap percentile confidence intervals of the means.

FIGURE 5. The number of loci (a), SNPs (b) and parsimony informative SNPs (PIS) (c) in relation to node depth. Observations are indicated with points. The number of PIS per taxon was logarithmically transformed as ln([number of PIS] + 1), one added because data include zeros, to ensure model goodness-of fit. The fitted regression curves (thick lines) and their 95% confidence limits (thin lines) are depicted, the regression equations being (a) $Y = 148 – 180X$ ($R^2 = 0.16$), (b) $Y = –101 + 7116X – 8239X^2$ ($R^2 = 0.30$) and (c) $Y = -0.513 + 9.12X$ ($R^2 = 0.48$); Y refers to the response variable and X to node depth.

FIGURE 6. Contour plots of the fitted regression surfaces explaining variation in bootstrap residuals in relation to node depth and either the number of loci (a), SNPs (b) or parsimony informative SNPs (c). The color gradient illustrates the shape of the regression surface, predicted negative and positive bootstrap residuals being indicated by blue and red colors, respectively. Observations are indicated with points, the color of the point being the darker the higher the bootstrap residual.