SCIENTIFIC REPORTS

OPEN

Received: 8 May 2017
Accepted: 31 July 2017
Published online: 29 August 2017

Reference-based RADseq resolves robust relationships among closely related species of lichen-forming fungi using metagenomic DNA

Felix Grewe[1], Jen-Pen Huang[1], Steven D. Leavitt[1,2] & H. Thorsten Lumbsch[1]

Despite increasing availability of phylogenomic datasets, strategies to generate genome-scale data from organisms involved in symbiotic relationships remains challenging. Restriction site-associated DNA sequencing (RADseq) can effectively generated reduced representation genomic loci. However, when using metagenomic DNA from inseparable symbiotic organisms, RADseq loci may belong to any number of the organisms involved in these intimate associations. In this study, we explored the potential for a reference-based RADseq approach to generate data for lichen-forming fungi from metagenomic DNA extracted from intact lichens. We simulated RAD data from draft genomes of closely related lichenized fungi to test if RADseq can reconstruct robust evolutionary relationships. Subsequently, we generated empirical RADseq data from metagenomic lichen DNA, with RADseq loci mapped back to a reference genome to exclude loci from other lichen symbionts that are represented in metagenomic libraries. In all cases, phylogenetic reconstructions using RADseq loci recovered diversification histories consistent with a previous study based on more comprehensive genome sampling. Furthermore, RADseq loci were found to resolve relationships among closely related species, which were otherwise indistinguishable using a phylogenetic species recognition criterion. Our studies revealed that a modified, reference-based RADseq approach can successfully be implemented to generate symbiont-specific phylogenomic data from metagenomic reads.

The continued development of high-throughput sequencing approaches enables time and cost efficient generation of massive amounts of genomic data and has facilitated the expansion of genomic datasets for phylogenetic- and population-level studies from single- and multi-locus matrices to genetic data sampled across entire genomes[1]. A wide range of methods are commonly used to generate genome-wide datasets, but rather than focusing on sequencing complete genomes, most methods are designed to sample subsets of organisms' genomes. Direct sequencing approaches, e.g. RNAseq[2,3] or genome skimming of highly abundant genomic regions or organelle genomes[4,5], and targeted approaches using baits to either pull genes of interest[6,7] or target ultra-conserved elements[8], can effectively generate reduced representation genome-scale datasets.

Among the most efficient reduced representation sequencing methods is restriction associated DNA sequencing (RADseq). In the original RADseq protocol, restriction enzymes were used to digest genomic DNA, and the resulting fragments are then sheared to sizes appropriate for a high-throughput sequencing[9,10]. Today, a plethora of RADseq techniques exist that vary from this original protocol, all of which share the common feature of sequencing loci flanking conserved restriction enzyme recognition motifs[11]. This flexibility allows RADseq protocols to be adjusted to best fit different genome sizes by the choice and number of restriction enzymes, thereby fragmenting genomic DNA to various degrees[12]. In addition, most protocols add a size-selection step to reduce the number of fragments that are sequenced so that a sufficient depth of coverage is obtained. While some initial knowledge about the targeted genome can be helpful (e.g. size and G/C content to determine the most appropriate restriction enzyme), RADseq can be applied without knowledge of a reference genome, hence is an ideal method for non-model organisms.

[1]Integrative Research Center, Science and Education, Field Museum of Natural History, 1400S Lake Shore Drive, Chicago, IL, 60605, USA. [2]Department of Biology & M. L. Bean Life Science Museum, Brigham Young University, Provo, UT, 84602, USA. Correspondence and requests for materials should be addressed to F.G. (email: fgrewe@fieldmuseum.org)

Based on its relatively uncomplicated setup and low costs, RADseq has been applied to the study of a wide range of organisms. RADseq has been used for both model and non-model organisms, and has answered a wide variety of biogeographical, ecological, evolutionary, and conservation-related questions[13–20]. RADseq has also been shown to successfully delimit species and reconstruct phylogenies of both recently diverged groups[21–24] and clades with diversifications histories spanning tens of millions of years[25–27].

Although RADseq is increasingly popular since its development, with over 600 publications mentioning any of the RADseq variations in the year 2016[11], this approach has rarely been used to study organisms involved in intimate, largely inseparable symbiotic relationships. Implementing RADseq protocols for these symbiotic organisms is challenging since the sequencing pool harbors metagenomic DNA from individuals from evolutionarily distinct lineages. A limited number of studies of symbiotic organisms using RADseq have circumvented this problem by either growing the targeted organism in axenic cultures or physically separating symbiotic partners by hand[28, 29]. However, for many organisms growing in intimate symbiotic associations, such as lichen-forming fungi, it is difficult to physically separate the symbiotic partners and growth of axenic cultures may be prohibitively slow[30, 31]. In these cases, a reference-guided approach is crucial to separate the metagenomic pool of sequences that was derived from all symbiotic partners of the holobiont.

Lichens provide interesting systems for studying diversification in obligate symbiotic systems[32–34]. Developing time and cost-effective approaches for generating genome-scale datasets is crucial to infer robust diversification histories and gain novel insight into evolutionary processes in these symbiotic fungi. However, to-date only a limited number of studies generated genome-scale datasets for inferring phylogenetic relationships[35, 36]. A reference-guided whole-genome skimming approach was used to generate various phylogenomic datasets for a clade of closely-related lichen-forming fungal species by mapping raw metagenomic sequence reads to a reference genome sequenced from an axenic fungal culture[36]. Although this study was able to infer robust phylogenetic relationships, the genome skimming approach was relatively costly and limited to few dozen individuals. It remains unclear how other reduced representation approaches for generating phylogenomic datasets can effectively be applied to metagenomic DNA pools isolated from lichen holobionts.

The *Rhizoplaca melanophthalma* species complex provides a useful study system to investigate if a reference-guided RADseq approach can successfully be implemented to generate meaningful phylogenomic data for the fungal partner from metagenomic lichen DNA. This species complex is comprised of nine species – *R. haydenii*, *R. idahoensis*, *R. melanophthalma*, *R. novomexicana*, *R. occulta*, *R. parilis*, *R. polymorpha*, *R. porteri*, and *R. shushanii*[37], which diversified largely during the Pliocene and into the Pleistocene[38]. This clade of closely related lichen-forming fungi is relatively well-studied[39], and crucial genomic and phylogenomic resources are available for comparisons, including a draft genome assembly based on an axenic fungal culture and other genome-scale datasets[36]. A number of species in this complex have broad, intercontinental distributions, and genome-scale data will be crucial to better understand population structure and dispersal capacity of these cosmopolitan species. In contrast, other species, including western North American endemics *R. haydenii* and *R. idahoensis*, have geographically restricted distributions and populations are threatened by habitat alteration due to agriculture, grazing, and invasive species[40]. Finally, relationships and boundaries among a number of closely related species, *R. occulta*, *R. polymorpha*, and *R. porteri* – the 'porteri group', remain unresolved[36].

In this study, we investigated the utility of RADseq for generating symbiont-specific phylogenomic datasets from metagenomic samples, such as symbiotic lichens. RADseq loci from lichen metagenomes were sorted by a mapping approach against a reference lichen-fungus genome sequence. We first tested if RADseq data can reconstruct evolutionary relationships within the *Rhizoplaca melanophthalma* multi-species complex consistent with those inferred from other, more comprehensive genome-wide datasets. For this study, we simulated RAD data from draft genome sequences from an earlier study. After successfully reconstructing a *Rhizoplaca* phylogeny from the simulated data and consistent with previous reconstructions, we generated empirical RADseq data from metagenomic lichen DNA and added the empirical RADseq data to the simulated data for a hybrid analysis. Finally, we investigated if the RADseq data can also be utilized for population genomics analysis to resolve shallow relationships in the *Rhizoplaca* phylogeny, investigating relationships in the 'porteri group'. Our studies revealed that a modified RADseq approach can successfully be implemented to generate symbiont-specific phylogenomic data from metagenomic reads in a cost- and time-efficient method.

Material and Methods

Genomic data used in RADseq simulation.

For the RADseq simulation approach, we utilized draft genome sequences from 30 specimens of the *R. melanophthalma* species complex from a previously published study and a reference genome generated from an axenic mycobiont culture of the lichenized fungi *Rhizoplaca melanophthalma sensu stricto*[36]. The taxon sampling for the RADseq simulation included: *R. haydenii* (n = 2), *R. melanophthalma* (6), *R. novomexicana* (1), *R. occulta* (2), *R. parilis* (4), *R. polymorpha* (6), *R. porteri* (5), and *R. shushanii* (5).

RADseq simulation.

Since lichens are symbiotic associations consisting of evolutionarily independent lineages and DNA extracted from a lichen thallus is comprised of genomes from all associated symbionts, we separated the metagenomic data of each sample with a mapping approach to a reference fungal genome from a *R. melanophthalma* culture. In Leavitt *et al.*[36], the reference genome sequence of this lichen fungus was *de novo* assembled by using the RAY v2.3.1 assembler[41, 42] with a kmer value of 41, and included all scaffolds with sizes larger than 5 kb for subsequent analyses. The total DNA of the symbiont genomes of the remaining taxa were sequenced with a low coverage, which allowed sorting of the reads by mapping them onto the reference sequence. Within the RealPhy v1.10 pipeline[43], raw reads of the low-covered 30 taxa were mapped to the reference genome with Bowtie2 v2.1.0[44]. RealPhy was used with the following arguments: -readLength 75 -perBaseCov 5 -gapThreshold 0.2. An average of 70.8% of the reference genome was covered by the mapping the metagenomes. A

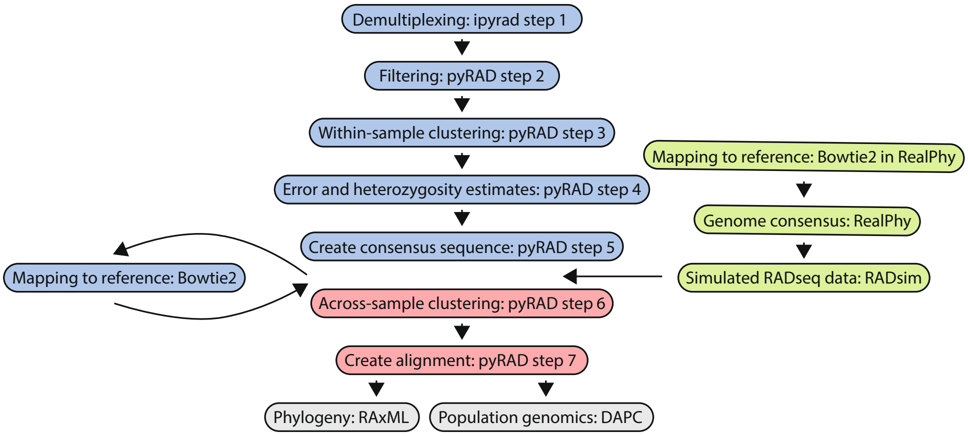

Demultiplexing: ipyrad step 1
Filtering: pyRAD step 2
Within-sample clustering: pyRAD step 3
Error and heterozygosity estimates: pyRAD step 4
Create consensus sequence: pyRAD step 5
Mapping to reference: Bowtie2
Across-sample clustering: pyRAD step 6
Create alignment: pyRAD step 7
Phylogeny: RAxML
Population genomics: DAPC
Mapping to reference: Bowtie2 in RealPhy
Genome consensus: RealPhy
Simulated RADseq data: RADsim

Figure 1. Computational workflow of analysis. Green boxes in the flow chart show the treatment of the simulated data before adding it to pyRAD (steps 6 and 7, in red) followed by a phylogenetic analysis (grey). Blue boxes in the flow chart represent the initial processing of the empirical data in ipyrad and pyRAD with an additional step of lichen fungus identification by mapping the metagenomic loci to the reference sequence. The mapped loci were then added to the simulated data (green) for a combined processing in pyRAD (steps 6 and 7, in red) prior to phylogenetic and population genomics analyses (grey).

consensus sequence of the successfully mapped reads was extracted for each specimen and used for the RADseq dataset simulation.

RADseq datasets of the *de novo* reference genome of *R. melanophthalma sensu stricto* and the 30 consensus sequences of the *R. melanophthalma* species complex were simulated by using the R package SimRAD[45]. After an initial test using different cut sites on the reference sequence to select for the best restriction enzyme, we *in silico* digested all 30 genome sequences with ApeKI (5-cutter with one ambiguous position: G|CWGC). The digested genome was subsequently selected for fragments with sizes between 200 and 500 bp. From these size-selected fragments, the first and last 143 bp sequences were extracted simulating a 150 bp Illumina sequencing after adapter and barcode sequences were removed. Since loci that are overlapping by more than 30% are merged in the pyRAD pipeline, we extracted the full sequence of the size-selected fragments instead of two separate sequences when the fragments were shorter than 243 bp.

The output was formatted to fit into the pyRAD v3.0.64[46] pipeline after step 5 when raw reads have been demultiplexed (step 1), quality filtered (step 2), within-sample clustered (step3), error and heterozygosity estimated (step 4), and have had consensus sequences (loci) created (step 5) (Fig. 1). The simulated data – representing the consensus sequences of each cluster (loci) – was across sample clustered in the pyRAD pipeline with vsearch[47] set to a 90% clustering threshold (Wclust = .90) (step 6). Finally, all clusters were filtered for a final alignment (step 7) that required a minimum of four samples for each final locus (MinCov = 4). The data type was set to 'gbs' since all fragments have an ApeKI cut site on both ends, which would qualify for sequencing from either direction. The resulting alignments of unlinked loci (.unlinked_snps) were used for phylogenetic analysis.

Genomic data used for empirical RADseq in the laboratory.

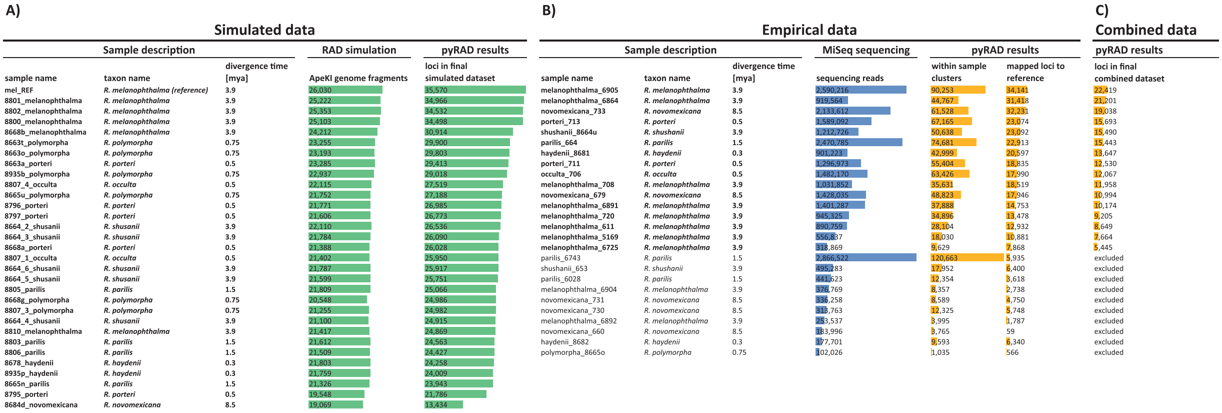
For an empirical comparison using genomic data generated by RADseq lab procedures, we selected 57 additional specimens of the *Rhizoplaca melanophthalma* species complex. All specimens were collected from sites throughout western North America (Supplementary table 1). Total metagenomic DNA was extracted either by following a modified CTAB protocol[48] or by using the Prepease DNA Isolation Kit (USB – product discontinued). Prior to the preparation of the RADseq libraries, we verified the identity of each specimen by sequencing their nuclear ITS rDNA region with a combination of primers ITS1f[49] and ITS4[50] and a comparison to a previously published worldwide sampling[37].

RADseq library production and sequencing in the laboratory.

RADseq libraries were prepared as described earlier[51] with an additional size selection step. In short, for the RADseq library production 100 ng of each DNA isolation was dried overnight together with approximately 0.06 pmol of adapters (designed as outlined elsewhere[52]), then resuspended for a digestion with the restriction enzyme ApeKI (New England Biolabs). This resuspension was followed by a ligation using a T4 ligase (New England Biolabs) as described[51]. Up to 48 samples with compatible barcodes were pooled and selected for fragments of sizes between 300 and 500 bp using the BluePippin (Sage Science). The pooled libraries were amplified using the REDTaq ReadyMix (Sigma-Aldrich) with primer pairs that were binding the ligated adapters (see ref. 52 for sequence details). The completed libraries were directly sequenced on an Illumina MiSeq using the MiSeq Reagent Kit v3 for 150 cycles (Illumina) to produce single-end sequences with a length of 150 bp.

Combined analysis of simulated and empirical RADseq datasets.

The raw reads from two MiSeq runs were individually demultiplexed in step one of the ipyrad workflow (https://github.com/dereneaton/ipyrad/blob/master/docs/index.rst) (Fig. 1). Subsequently, steps two to five were processed with pyRAD v3.0.64[46] for genotype-by-sequencing data (Datatype = gbs) of haploid genomes (ploidy = 1). Within-sample clustering was done by vsearch[47] set to a 90% clustering threshold (Wclust = 0.90). The consensus sequence was generated from



A)

Simulated data

Sample description			RAD simulation	pyRAD results
sample name	taxon name	divergence time [mya]	ApeKI genome fragments	loci in final simulated dataset
mel_REF	***R. melanophthalma (reference)***	**3.9**	26,030	35,570
8801_melanophthalma	***R. melanophthalma***	**3.9**	25,222	34,966
8802_melanophthalma	***R. melanophthalma***	**3.9**	25,353	34,532
8800_melanophthalma	***R. melanophthalma***	**3.9**	25,103	34,498
8668b_melanophthalma	***R. melanophthalma***	**3.9**	24,212	30,914
8663t_polymorpha	***R. polymorpha***	**0.75**	23,255	29,900
8663o_polymorpha	***R. polymorpha***	**0.75**	23,193	29,803
8663a_porteri	***R. porteri***	**0.5**	23,285	29,413
8935b_polymorpha	***R. polymorpha***	**0.75**	22,937	29,018
8807_4_occulta	***R. occulta***	**0.5**	22,115	27,519
8665u_polymorpha	***R. polymorpha***	**0.75**	21,752	27,188
8796_porteri	***R. porteri***	**0.5**	21,771	26,985
8797_porteri	***R. porteri***	**0.5**	21,606	26,773
8664_2_shusanii	***R. shusanii***	**3.9**	22,110	26,536
8664_3_shusanii	***R. shusanii***	**3.9**	21,784	26,090
8668a_porteri	***R. porteri***	**0.5**	21,388	26,028
8807_1_occulta	***R. occulta***	**0.5**	21,402	25,950
8664_6_shusanii	***R. shusanii***	**3.9**	21,787	25,917
8664_5_shusanii	***R. shusanii***	**3.9**	21,599	25,751
8805_parilis	***R. parilis***	**1.5**	21,809	25,066
8668g_polymorpha	***R. polymorpha***	**0.75**	20,548	24,986
8807_3_polymorpha	***R. polymorpha***	**0.75**	21,255	24,982
8664_4_shusanii	***R. shusanii***	**3.9**	21,100	24,915
8810_melanophthalma	***R. melanophthalma***	**3.9**	21,417	24,869
8803_parilis	***R. parilis***	**1.5**	21,612	24,563
8806_parilis	***R. parilis***	**1.5**	21,509	24,427
8678_haydenii	***R. haydenii***	**0.3**	21,803	24,258
8935p_haydenii	***R. haydenii***	**0.3**	21,759	24,009
8665n_parilis	***R. parilis***	**1.5**	21,326	23,943
8795_porteri	***R. porteri***	**0.5**	19,548	21,786
8684d_novomexicana	***R. novomexicana***	**8.5**	19,069	13,434

B)

Empirical data

Sample description			MiSeq sequencing	pyRAD results	
sample name	taxon name	divergence time [mya]	sequencing reads	within sample clusters	mapped loci to reference
melanophthalma_6905	***R. melanophthalma***	**3.9**	2,590,216	90,253	34,141
melanophthalma_6864	***R. melanophthalma***	**3.9**	919,564	44,767	31,418
novomexicana_733	***R. novomexicana***	**8.5**	2,133,612	61,528	32,231
porteri_713	***R. porteri***	**0.5**	1,589,092	67,165	23,074
shushanii_8664u	***R. shushanii***	**3.9**	1,212,726	50,638	23,092
parilis_664	***R. parilis***	**1.5**	2,470,785	74,681	22,913
haydenii_8681	***R. haydenii***	**0.3**	901,223	42,999	20,597
porteri_711	***R. porteri***	**0.5**	1,296,973	55,404	18,835
occulta_706	***R. occulta***	**0.5**	1,482,170	63,426	17,990
melanophthalma_708	***R. melanophthalma***	**3.9**	1,031,852	35,631	18,519
novomexicana_679	***R. novomexicana***	**8.5**	1,428,035	48,823	17,946
melanophthalma_6891	***R. melanophthalma***	**3.9**	1,401,287	37,888	14,753
melanophthalma_720	***R. melanophthalma***	**3.9**	945,325	34,896	13,478
melanophthalma_611	***R. melanophthalma***	**3.9**	890,759	28,104	12,932
melanophthalma_5169	***R. melanophthalma***	**3.9**	556,837	18,030	10,881
melanophthalma_6725	***R. melanophthalma***	**3.9**	318,869	9,629	7,868
parilis_6743	*R. parilis*	1.5	2,866,522	120,663	5,935
shushanii_653	*R. shushanii*	3.9	495,283	17,952	6,400
parilis_6028	*R. parilis*	1.5	441,623	12,354	3,618
melanophthalma_6904	*R. melanophthalma*	3.9	376,769	8,357	2,738
novomexicana_731	*R. novomexicana*	8.5	336,258	8,589	4,750
novomexicana_730	*R. novomexicana*	8.5	313,763	12,325	5,748
melanophthalma_6892	*R. melanophthalma*	3.9	253,537	3,995	1,787
novomexicana_660	*R. novomexicana*	8.5	183,996	3,765	59
haydenii_8682	*R. haydenii*	0.3	177,701	9,593	6,340
polymorpha_8665o	*R. polymorpha*	0.75	102,026	1,035	566

C)

Combined data

pyRAD results
loci in final combined dataset
22,419
21,201
19,038
15,693
15,490
15,443
13,647
12,530
12,067
11,958
10,994
10,174
9,205
8,649
7,664
5,445
excluded
excluded
excluded
excluded
excluded
excluded
excluded
excluded
excluded
excluded

Figure 2. Overview of the RADseq results after individual steps of analyses. (**A**) Number of fragments and loci in the final dataset of the simulation. (**B**) Number of raw sequences, within sample clusters (after pyRAD step 5), and mapped loci to the reference sequence of the empirical dataset. (**C**) Number of loci in the final combined dataset (after final pyRAD step 7). Bars represent the respective values in relation to their colored group. Divergence times were used according to Leavitt *et al.*[38]. Additional 31 samples with less than 20,000 sequences were excluded (not shown in table). Samples in bold had a total number of more than 5,000 loci in the final dataset and were included for further analyses.

all clusters with a minimum coverage of four reads (Mindepth = 4). These consensus sequences – representing all the recovered metagenomic loci from each sample – were selected for sequences of lichen fungal origin by mapping them to the reference sequence with Bowtie2 with adjusted parameters to one permitted mismatch (−N 1), a seed length of 20 (−L 20), up to 20 seed extension attempts (−D 20), and a maximum "re-seeding" of 3 (−R 3). The mapped loci were then combined with the simulated data for an across sample clustering in pyRAD (step 6 and step 7) with the same parameter setting as used for only simulated data (see above; Wclust = 0.90; MinCov = 4). Samples for which the final statistics output indicated less than 5000 loci were removed for a repeated run of step 6 and step 7. In an additional analysis, we lowered the missing data in the final alignment by increasing the minimum number of samples per final loci (MinCov = 30). Phylogenetic analyses were done with the resulting alignments that only included the unlinked loci (.unlinked_snps) of the combined data from empirical and simulated samples.

Correlation of data from samples of simulated and empirical origin was tested by a linear regression model in R[53]. The proportion of shared loci of the combined simulated and empirical data were generated in a pairwise similarity matrix and displayed by using the R package RADami[25].

Phylogenetic Analysis of RADseq datasets. Phylogenetic trees were estimated using maximum likelihood interference with RAxML v7.2.8[54] using the GTR + G + I model. For each analysis, 100 bootstrap replicates were calculated using the fast bootstrapping option implemented in RAxML[55]. Trees were drawn with the TreeExplorer implemented in MEGA 7.0.20[56]. Based on previous studies, all resulting trees were rooted with *R. novomexicana*.

The genetic structure of the members of the 'porteri group' was evaluated with the Discriminant Analysis of Principal Components (DAPC) implemented in the R package adegenet v2.0.1[57, 58]. This method relies on a data transformation by a Principal Component Analysis (PCA) prior to the separation of individuals of a population by their genetic distance using a Discriminant Analysis (DA). To minimize the missing data for the PCA, we treated the dataset by using the "mean" option for genotypic data type as recommended[57] and excluded loci that had more than 20% missing data. For the DAPC, the data was loaded into an R data frame with settings for a haploid genome (ploidy = 1). The DAPC was then conducted by using the first five principal components and all (two) DA-eigenvalues. In addition to the display of the genetic variation in the genomic space, the DAPC allows a prediction of the group membership probability for each sample. Group memberships for samples were predefined according to the corresponding 'taxon name' in Fig. 2.

Reproducibility. All scripts that were used in this study are available (https://github.com/felixgrewe/lichen_RADseq). All RAD sequences were deposited in the NCBI Sequence Read Archive (SRA) with accession numbers SRR5807616 - SRR5807612.

Results and Discussion

RADseq simulation. For the RADseq simulation, the reference genome sequence of *R. melanophthalma* was *in silico* digested by six different restriction enzymes (SbfI, PstI, NsiL, BclI, BstYI, and ApeKI). The resulting fragments of sizes between 200 and 500 bp were counted to identify the best restriction enzyme for this study (Supplementary Fig. 1). The lichen-fungal genome of *R. melanophthalma* has a small size of 38.7 Mb[36], which demands the use of a common-cutting restriction enzyme that produces a sufficient number of fragments for the RADseq analysis. The simulated use of 6- and 7-cutter restriction enzymes produced an insufficient number of fragments for RADseq (0–1,173 fragments, 3,463 fragments when the recognition motif of the 6-cutter

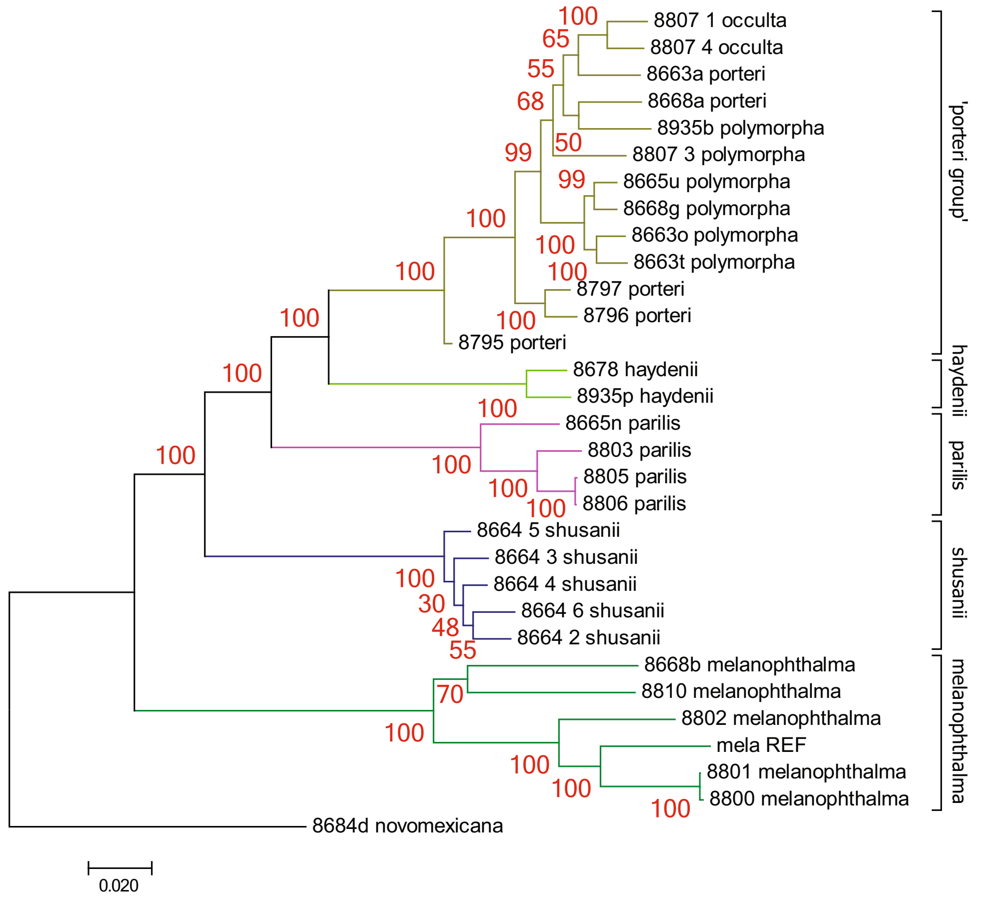


Figure 3. Phylogenetic tree inferred from the simulated RAD data. Samples representing each species are highlighted with individual branch colors. Bootstrap values are indicated near nodes in red. The unit of the branch lengths is substitutions per site.

had two ambiguous sites). However, we highly increased the fragment number to 26,030 by cutting the genome with ApeKI – a 5-cutter with one ambiguous site in its recognition motif. We then selected ApeKI for all further analyses, since it produced enough fragments to represent sufficient genomic variation of the relatively small lichen-forming fungal genomes – ca. 40 Mb – and compensate for a potential loss of homologous loci due to allele drop-out (the loss of loci due to mutations in the enzyme recognition motif.)

Between 19,069 and 25,222 fragments were produced using the restriction enzyme ApeKI on each of the *Rhizoplaca* genomes in the simulation (Fig. 2A). These fragments were integrated into the pyRAD pipeline (after step 5, Fig. 1) and across sample clustered into 13,434 to 35,570 loci per sample. The final alignment contained 53,659 positions and 56.77% gaps.

The phylogeny inferred from the simulated RAD loci (Fig. 3), representing a sub-sampling of the genomic data to a reduced genome representation, recovered identical phylogenetic relationships among species compared to the most comprehensive phylogenomic dataset – 'RealPhy' – in Leavitt *et al.*[36]. All species of the *Rhizoplaca melanophthalma* species complex were recovered as individual monophyletic groups supported by strong bootstrap support except for the 'porteri group', which included *R. polymorpha*, *R. porteri*, and *R. occulta*. To extend the current phylogeny, we generated further RAD loci from more individuals in the laboratory and incorporated them into the simulated dataset for a combined processing of simulated and empirical RAD loci using pyRAD.

RAD sequencing. For the RAD library production in the laboratory, we built two libraries of total lichen metagenomic DNA isolations. These libraries were sequenced by a total of 17,7 × 10^6 and 19,8 × 10^6 reads, respectively. The number of reads of each successfully sequenced sample (>20,000 reads) varied widely, from 102,026 to 2,8 × 10^6 reads with an average of 1,03 × 10^6 (sd = 794,051) reads per sample (Fig. 2B). This variation is most likely caused by an uneven normalization of the libraries prior to sequencing. A proper normalization of the libraries is an important step which avoids overrepresentation of a few samples and warrants maximum efficiency from the sequencing. The number of within sample clusters that pyRAD generated from these sequences directly correlated with the initial number of sequences ($R^2 = 0.9295$, not shown) indicating that the higher the initial sequencing, the more clusters were generated.

A crucial step in this pipeline was the identification of the fungal clusters from the pool of metagenomic within-sample clusters by mapping them to the lichen-fungus genome of *R. melanophthalma*. The mapping detected that an average of 43% (sd = 18%) of the clusters originated from the fungal genome leaving between

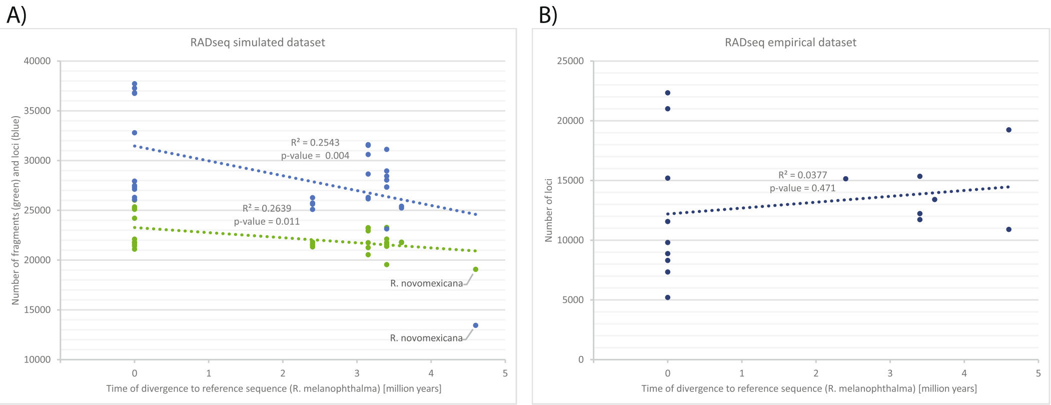


Figure 4. Correlation of RADseq results to the phylogenetic distance of the samples to the reference sequence (*R. melanophthalma*). (**A**) The number of fragments and loci of the final simulated dataset are shown in green and blue, respectively. (**B**) The number of loci of the empirical RADseq samples are indicated in orange. All numbers are plotted against the time of divergence of samples to the reference, *R. melanophthalma*. Trendlines are based on the linear regression model and are drawn in the respective colors of the different datasets.

59 to 34,141 loci for each sample in the across-sample clustering step (Fig. 2B). Therefore, in most metagenomic pools of samples the number of lichen-fungal loci were smaller than the number of other loci from an unidentified origin. Multiple studies reported that lichens consist of more than just mycobiont and photobiont genomes, but can also harbor multiple algae species[59–61], high content of bacteria[62–65], and/or even additional, distantly related, fungi[32]. Any of these organisms could have contributed to the metagenomic pool and increased the amount of non-lichen fungal sequences. With an increasing number of potential reference genomes for these organisms available in Genbank, the identification of these loci of unknown origin could be integrated into interesting future studies.

Two samples differed from the general observation that around half of the metagenomic pool has fungal origin. The *R. parilis* sample 'parilis_6743' had most raw reads (2,9 × 10^6) and most metagenomic clusters (120,663) of all samples, but only 5,935 of these clusters (5%) consisted of fungal DNA. Since this pattern was not consistent throughout all the *R. parilis* samples, these numbers may reflect contamination or misidentification of the sample. In contrast, the *R. melanophthalma* sample 'melanophthalma_6725' contained over 82% fungal clusters. Hence, it could be included in the final dataset although the sample only had a relatively small number of initial raw sequences.

On average, 67% (sd = 2.9%) of the fungal loci successfully across-sample clustered with loci from at least three other samples (Fig. 2C). After this clustering, a total of 47 samples remained with more than 5,000 loci. These samples were included in the final dataset of 76,371 loci and a final alignment of 67,775 positions with 70.46% gaps.

Correlation of simulated and empirical RAD data.

For the simulated dataset, reads were initially mapped to the reference to create a consensus sequence of the lichen-fungal genome for each sample. The number of RAD fragments and loci that were recovered from lichen metagenomes was dependent on the distance of the relationship between samples and the reference genome that was targeted (Fig. 4A). The relationship of the species had less effect on the mapping, represented here by the number of fragments after a simulated digest ($R^2 = 0.2639$, $p = 0.011$), rather than on the number of final loci produced by the pyRAD pipeline ($R^2 = 0.2543$, $p = 0.004$). This correlation also remained significant after excluding the outgroup *R. novomexicana* for the fragments ($R^2 = 0.2093$, $p = 0.011$) and loci ($R^2 = 0.1805$, $p = 0.019$). For *R. novomexicana*, which is most distantly related to the reference, the number of loci dropped much lower than projected by the regression model. The drop of loci cannot be caused by allele dropout since the simulated digest produced fragment numbers close to the prediction. However, it demonstrates that the mapping of the raw reads is less sensitive (allows more mapping) than the cluster generation in pyRAD under the current parameter settings. The processing of more distantly related taxa would therefore require an adjustment of the clustering threshold, which also increases the risk of clustering paralogous and non-homologous loci. The mapping approach implemented in 'RealPhy' may be less effected by genomic variation and therefore suitable to recover more loci from distantly related taxa. However, more distantly related individuals would be needed in this study to further support this observation.

For the empirical dataset, reads were first within sample clustered in pyRAD before they were sorted by mapping to the fungal reference sequence. The number of final loci of the empirical dataset is not correlated with an increasing distance of the species relationship (Fig. 4B) and differed from the projection based on the simulated dataset (Fig. 4A). This difference indicates that the variation in loci number results from biases during the RADseq laboratory procedures rather than from errors in the subsequent computational sequence analysis. A lower number of loci also leads to fewer shared loci across the empirical samples compared to simulated samples as indicated in a pairwise comparison of the data (Fig. 5).

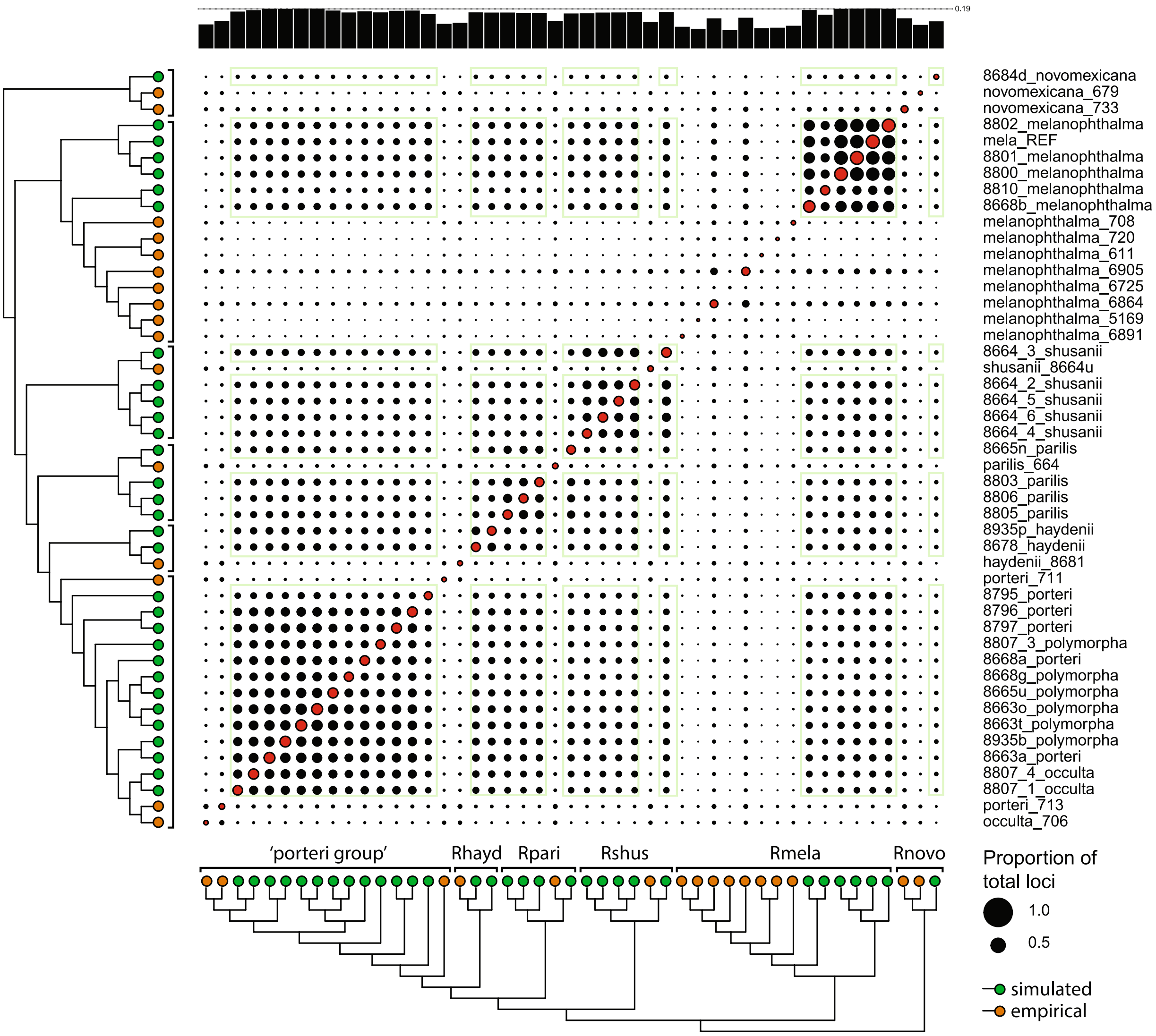


Figure 5. Proportion of shared loci among individuals of simulated and empirical origin. Correlation plot shows percentage of shared loci and successfully amplified loci for each sample in black and red circles, respectively. Green boxes comprise correlations of samples derived from simulated data. Black bars on top of the plot indicate the average percentage of shared loci per sample. The phylogenetic trees were calculated from the combined dataset as in Fig. 5 with green and orange tips indicating samples derived from simulated and empirical data, respectively.

Phylogenetic analysis of combined data. Phylogenetic analyses of the combined dataset recovered all species of the *Rhizoplaca* species complex as monophyletic groups, except for those within the 'porteri group', in which the species *R. occulta*, *R. polymorpha*, and *R. porteri* were recovered intermixed in a single, well-supported clade (Fig. 6A). The tree topology is therefore similar to inferences from the simulated RAD data (Fig. 3) and the whole-genome dataset used in Leavitt *et al.*[36]. In this topology, all newly included 16 samples clustered within their respective species groups, and the monophyly of each group remained strongly supported, with the exception of those in the 'porteri group'. This group remained unresolved even after the addition of new samples. Moreover, the support for the monophyly of the 'porteri group' decreased (bootstrap = 72) largely due to the addition of specimen 'porteri_711', which was resolved as sister to all remaining group taxa. Using a smaller alignment (at least 30 samples/loci, 7770 positions, 27.1% gaps), specimen 'porteri_711' was recovered as sister to the 'porteri group' and *R. haydenii*, which explains that the weaker bootstrap support in both trees is due to its unstable position (Supplementary Fig. 2). After including new samples, another difference occurred within the 'porteri group' where the added specimen 'porteri_713' clustered together with the remaining *R. occulta* individuals and disrupted the former monophyly of the group.

Since the 'porteri group' remained unresolved by the maximum likelihood tree reconstruction, we tested population genetic approaches using the RAD data generated here to differentiate the genomes by their variation. In preparation for the DAPC, we filtered the dataset on missing data which resulted in a total of 4997 qualified loci. The DAPC combines a PCA with a DA that supports a separation of genomes based on their variance between

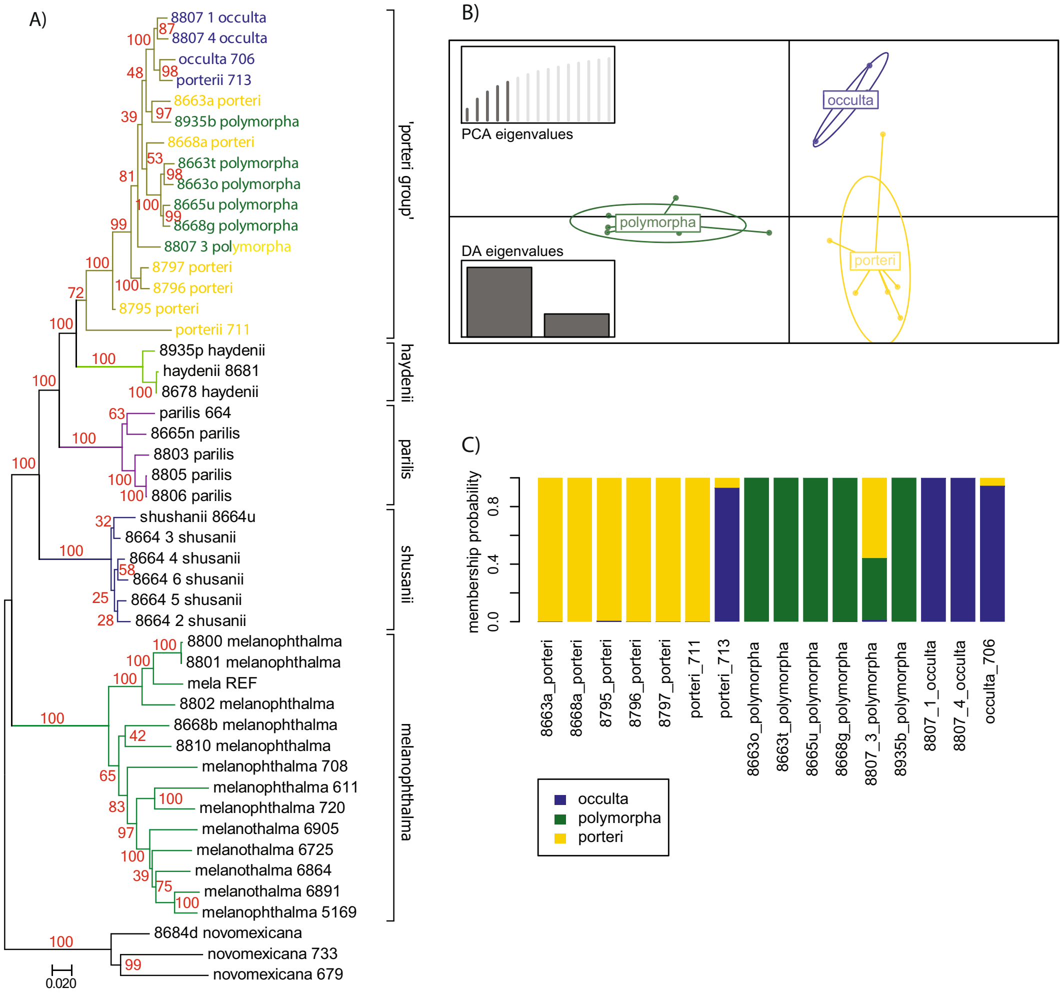


Figure 6. Relationship of the *R. melanophthalma* species complex estimated from the combined dataset. (**A**) Phylogenetic tree inferred from the combined dataset. Samples representing each species are highlighted with branch colors as in Fig. 2. Bootstrap values are represented by red numbers near nodes. The unit of the branch lengths is substitutions per site. The color of the label of the samples of the 'porteri group' show the group membership probability estimated with the DAPC. (**B**) DAPC scatterplot of samples of the 'porteri group'. Individuals and groups are represented by dots and inertia ellipses, respectively. (**C**) STRUCTURE-like plot of group membership probabilities of samples of the 'porteri group'.

groups rather than their variance within groups. All three species of the 'porteri group' built clearly separated clusters supporting their genomic divergence from each other (Fig. 6B). In addition, the DAPC method exploits group memberships of individuals, which are represented by a STRUCTURE-like plot (Fig. 6C). In this analysis, all individuals of the 'porteri group' were assigned high probabilities to distinct groups corresponded to their nominal species, with the exception of specimen '8807_3_polymorpha', which the membership probabilities were divided between *R. polymorpha* and *R. porteri*. All other *R. polymorpha* and *R. occulta* specimens belonged to their respective membership groups. Of the samples representing *R. porteri*, however, one specimen, 'porteri_773', was assigned membership to the *R. occulta* group with high probability. The same sample also clustered with high support within the *R. occulta* clade in the phylogenetic tree (Fig. 6A). Hence, the DAPC method may have identified a misidentified sample, under which condition the *R. occulta* would remain monophyletic as earlier observed (Fig. 3, ref. [36]).

The combined RAD dataset was sufficient to infer well-resolved phylogenies. Even though some of the samples lacked variation and were not well separated in the phylogenetic trees, these samples could be distinctly separated by population genetic methods such as DAPC. Compared to other reduced genome representation methods such as sequence capturing, RADseq produces a higher number of loci with relatively little effort in the laboratory and much lower costs[66]. Therefore, future phylogenetic and population genetic analyses of lichen-fungal genomes may solely rely on reduced genome representation data, such as RADseq, which will dramatically reduce sequencing costs and allow a deeper (taxon-) sampling for each study.

Conclusion

We successfully implemented a strategy for generating RADseq loci for targeted lineages involved in intimate symbiotic associations, lichen-forming fungi. Whole lichens were sequenced and reduced representation RADseq metagenomic libraries were filtered for loci derived from the lichen-fungal genome using a reference-guided mapping approach. The resulting reduced genome representation dataset had sufficient phylogenetic signal to reconstruct phylogenetic trees consistent with those using more comprehensive genome-scale datasets. A limiting factor of this reference-guided RADseq method is the requirement of a closely related reference genome. This requirement, however, will be less of a restraining factor in the future with an increasing number of complete lichenized fungal genomes available in Genbank. In addition, applications for metagenomic high-throughput sequencing are increasingly sophisticated for separating *de novo* in-depth sequenced metagenomes. These advances will allow for future studies to create reliable lichen-fungus reference sequences to sort metagenomic RAD loci. Hence, the low-cost RADseq method presented here can be applied for reduced genome representation of other lichen species. Thereby the opportunity to sample thousands of individuals in just one high-throughput sequencing run will open new avenues for lichen phylogenomic and population genomics analyses.

References

1. Davey, J. *et al*. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510 (2011).
2. Wickett, N. J. *et al*. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci.* **111**, E4859–E4868 (2014).
3. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).
4. Greshake, B. *et al*. Potential and pitfalls of eukaryotic metagenome skimming: A test case for lichens. *Mol. Ecol. Resour.* **16**, 511–523 (2016).
5. Dodsworth, S. Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.* **20**, 525–527 (2015).
6. Weitemier, K. *et al*. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. *Appl. Plant Sci.* **2**, 1400042 (2014).
7. Johnson, M. G. *et al*. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. *Appl. Plant Sci.* **4**, 1600016 (2016).
8. Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C. & Brumfield, R. T. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Syst. Biol.* **63**, 83–95 (2014).
9. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, 1–7 (2008).
10. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. & Johnson, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* **17**, 240–248 (2007).
11. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. & Hohenlohe, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **17**, 81–92 (2016).
12. Herrera, S., Reyes-Herrera, P. H. & Shank, T. M. Predicting RAD-seq marker numbers across the eukaryotic tree of life. *Genome Biol. Evol.* **7**, 3207–3225 (2015).
13. Nadeau, N. J. *et al*. Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome Res.* **24**, 1316–33 (2014).
14. Cooper, E. A. & Uy, J. A. C. Genomic evidence for convergent evolution of a key trait underlying divergence in island birds. *Mol. Ecol.* **0**, 1–15 (2017).
15. Hoffman, J. I. *et al*. High-throughput sequencing reveals inbreeding depression in a natural population. *Proc. Natl. Acad. Sci.* **111**, 3775–3780 (2014).
16. Hohenlohe, P. A. *et al*. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Mol. Ecol.* **22**, 3002–3013 (2013).
17. Eaton, D. A. R., Hipp, A. L., Gonzalez-Rodriguez, A. & Cavender-Bares, J. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution (NY).* **69**, 2587–2601 (2015).
18. Derkarabetian, S., Burns, M., Starrett, J. & Hedin, M. Population genomic evidence for multiple Pliocene refugia in a montane-restricted harvestman (Arachnida, Opiliones, Sclerobunus robustus) from the southwestern United States. *Mol. Ecol.* **25**, 4611–4631 (2016).
19. Kolar, F. *et al*. Northern glacial refugia and altitudinal niche divergence shape genome-wide differentiation in the emerging plant model Arabidopsis arenosa. *Mol. Ecol.* **25**, 3929–3949 (2016).
20. Leache, A. D., Fujita, M. K., Minin, V. N. & Bouckaert, R. R. Species delimitation using genome-wide SNP Data. *Syst. Biol.* **63**, 534–542 (2014).
21. Winston, M. E. *et al*. Understanding cultivar-specificity and soil determinants of the Cannabis microbiome. *PLoS One* **9**, e99641 (2014).
22. Bell, R. C., Drewes, R. C. & Zamudio, K. R. Reed frog diversification in the Gulf of Guinea: Overseas dispersal, the progression rule, and *in situ* speciation. *Evolution (NY).* **69**, 904–915 (2015).
23. Bryson, R. W., Savary, W. E., Zellmer, A. J., Bury, R. B. & McCormack, J. E. Genomic data reveal ancient microendemism in forest scorpions across the California Floristic Province. *Mol. Ecol.* **25**, 3731–3751 (2016).
24. Mort, M. E. *et al*. Multiplexed-shotgun-genotyping data resolve phylogeny within a very recently derived insular lineage. *Am. J. Bot.* **102**, 634–641 (2015).
25. Hipp, A. L. *et al*. A Framework Phylogeny of the American Oak Clade Based on Sequenced RAD Data. **9** (2014).
26. Rubin, B. E. R., Ree, R. H. & Moreau, C. S. Inferring phylogenies from RAD sequence data. *PLoS One* **7**, 1–12 (2012).
27. Cariou, M., Duret, L. & Charlat, S. Is RAD-seq suitable for phylogenetic inference? An *in silico* assessment and optimization. *Ecol. Evol.* **3**, 846–852 (2013).
28. Grillo, M. A., De Mita, S., Burke, P. V., Solorzano-Lowell, K. L. S. & Heath, K. D. Intrapopulation genomics in a model mutualist: Population structure and candidate symbiosis genes under selection in Medicago truncatula. *Evolution (NY).* **70**, 2704–2717 (2016).
29. Wyss, T., Masclaux, F. G., Rosikiewicz, P., Pagni, M. & Sanders, I. R. Population genomics reveals that within-fungus polymorphism is common and maintained in populations of the mycorrhizal fungus Rhizophagus irregularis. *ISME J.* 1–13 (2016).
30. Crittenden, P. D. & Porter, N. Lichen-forming fungi: potential sources of novel metabolites. *Trends Biotechnol.* **9**, 409–14 (1991).
31. Crittenden, P. D., David, J. C., Hawksworth, D. L. & Campbell, F. S. Attempted isolation and success in the culturing of a broad spectrum of lichen-forming and lichenicolous fungi. *New Phytol.* **130**, 267–297 (1995).
32. Spribille, T. *et al*. Basidiomycete yeasts in the cortex of ascomycete macrolichens. *Science* **353**, 488–492 (2016).
33. Lutzoni, F., Pagel, M. & Reeb, V. Major fungal lineages are derived from lichen symbiotic ancestors. *Nature* **411**, 937–940 (2001).
34. Hawksworth, D. L. The variety of fungal-algal symbioses, their evolutionary significance, and the nature of lichens. *Bot. J. Linn. Soc.* **96**, 3–20 (1988).

35. Xavier, B. B., Miao, V. P. W., Jónsson, Z. O. & Andrésson, Ó. S. Mitochondrial genomes from the lichenized fungi Peltigera membranacea and Peltigera malacea: Features and phylogeny. *Fungal Biol.* **116**, 802–814 (2012).
36. Leavitt, S. D. *et al.* Resolving evolutionary relationships in lichen-forming fungi using diverse phylogenomic datasets and analytical approaches. *Sci. Rep.* **6**, 22262 (2016).
37. Leavitt, S. *et al.* DNA barcode identification of lichen-forming fungal species in the Rhizoplaca melanophthalma species-complex (Lecanorales, Lecanoraceae), including five new species. *MycoKeys* **7**, 1–22 (2013).
38. Leavitt, S. D. *et al.* Local representation of global diversity in a cosmopolitan lichen-forming fungal species complex (*Rhizoplaca*, Ascomycota). *J. Biogeogr.* **40**, 1792–1806 (2013).
39. Leavitt, S. D. *et al.* Complex patterns of speciation in cosmopolitan 'rock posy' lichens - Discovering and delimiting cryptic fungal species in the lichen-forming Rhizoplaca melanophthalma species-complex (Lecanoraceae, Ascomycota). *Mol. Phylogenet. Evol.* **59**, 587–602 (2011).
40. Rosentreter, R. Vagrant Lichens in North America. *Bryologist* **96**, 333 (1993).
41. Boisvert, S., Laviolette, F. & Corbeil, J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* **17**, 1519–1533 (2010).
42. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* **13**, R122 (2012).
43. Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B. & van Nimwegen, E. Automated reconstruction of whole-genome phylogenies from short-sequence reads. *Mol. Biol. Evol.* **31**, 1077–88 (2014).
44. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
45. Lepais, O. & Weir, J. T. SimRAD: An R package for simulation-based prediction of the number of loci expected in RADseq and similar genotyping by sequencing approaches. *Mol. Ecol. Resour.* **14**, 1314–1321 (2014).
46. Eaton, D. A. R. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844–1849 (2014).
47. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ Prepr.* **4**, e2409v1 (2016).
48. Cubero, O. F., Crespo, A., Fatehi, J. & Bridge, P. D. DNA extraction and PCR amplification method suitable for fresh, herbarium-stored, lichenized, and other fungi. *Plant Syst. Evol.* **216**, 243–249 (1999).
49. Gardes, M. & Bruns, T. D. ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Mol. Ecol.* **2**, 113–118 (1993).
50. White, T. J., Bruns, S., Lee, S. & Taylor, J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. *PCR Protocols: A Guide to Methods and Applications* 315–322 (1990).
51. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* **6**, 1–10 (2011).
52. Rubin, B. E. R. *et al.* Comparative genomics reveals convergent rates of evolution in ant–plant mutualisms. *Nat. Commun.* **7**, 12679 (2016).
53. R Development Core Team. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput. Vienna Austria* (2016).
54. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
55. Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–71 (2008).
56. Kumar, S. *et al.* MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
57. Jombart, T. *et al.* Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
58. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).
59. Catalá, S. *et al.* Coordinated ultrastructural and phylogenomic analyses shed light on the hidden phycobiont diversity of Trebouxia microalgae in Ramalina fraxinea. *Mol. Phylogenet. Evol.* **94**, 765–77 (2016).
60. Casano, L. M. *et al.* Two Trebouxia algae with different physiological performances are ever-present in lichen thalli of Ramalina farinacea. Coexistence versus Competition? *Environ. Microbiol.* **13**, 806–818 (2011).
61. Moya, P. *et al.* Unexpected associated microalgal diversity in the lichen Ramalina farinacea is uncovered by pyrosequencing analyses. *PLoS One* **12**, e0175091 (2017).
62. Suzuki, M. T., Parrot, D., Berg, G., Grube, M. & Tomasi, S. Lichens as natural sources of biotechnologically relevant bacteria. *Appl. Microbiol. Biotechnol.* **100**, 583–595 (2016).
63. Printzen, C., Fernández-Mendoza, F., Muggia, L., Berg, G. & Grube, M. Alphaproteobacterial communities in geographically distant populations of the lichen *Cetraria aculeata*. *FEMS Microbiol. Ecol.* **82**, 316–325 (2012).
64. Hodkinson, B. P. & Lutzoni, F. A microbiotic survey of lichen-associated bacteria reveals a new lineage from the Rhizobiales. *Symbiosis* **49**, 163–180 (2009).
65. Grube, M., Cardinale, M., de Castro, J. V., Müller, H. & Berg, G. Species-specific structural and functional diversity of bacterial communities in lichen symbioses. *ISME J.* **3**, 1105–1115 (2009).
66. Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C. & Brumfield, R. T. Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Syst. Biol.* **65**, 910–924 (2016).

Acknowledgements

The authors thank the Negaunee Foundation and the Field Museum for financial support, and the Lauer family for the generous gift of the MiSeq sequencer for the Pritzker Lab for Molecular Systematics of the Field Museum. We also thank Kevin Feldheim and Todd Widhelm (Chicago) for their invaluable help in the laboratory and Sarah Cowles (Miami) for comments on the manuscript. Financial support by the National Science Foundation (DEB-1354884) is acknowledged.

Author Contributions

F.G., S.D.L. and H.T.L. designed the research. F.G. and J.H. assembled molecular data and conducted phylogenetic analyses. F.G., J.H., S.D.L. and H.T.L. wrote the paper.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-09906-7

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.