Insect Systematics and Diversity, 1(2), 2017, 1–8
doi: 10.1093/isd/ixx009
Research Article

OXFORD

An Empirical Test of Reduced-Representation Genomics to Infer Species-Level Phylogenies for Two Ant Groups

Corrie S. Moreau[1] and Brian D. Wray

Field Museum of Natural History, Department of Science and Education, Integrative Research Center, 1400 South Lake Shore Drive, Chicago, IL 60605, and [1]Corresponding author, e-mail: cmoreau@fieldmuseum.org

Subject Editor: Kevin Johnson

Received 15 August 2017; Editorial decision 22 September 2017

Abstract

DNA sequence data generation has traditionally been a significant bottleneck in the production of well-resolved molecular phylogenies both in terms of time and money. As smaller laboratories now have access to molecular techniques once accessible only in large laboratories with expensive equipment, and the cost per base of DNA sequencing has dramatically dropped, most laboratories, including those working on nonmodel organisms, can produce large molecular datasets. In this study, we discuss the technical and financial details of producing a reduced-representation genomic dataset for the resolution of species-level phylogenies of two distantly related ant genera (*Cephalotes* and *Polyrhachis*), and then compare our resulting phylogenies with phylogenies generated by previous studies using the more traditional gene-based approach. We demonstrate that genotyping-by-sequencing is a cost-effective and appropriate method for species-level and potentially higher phylogenetics in insects but recognize that bioinformatic skills will now be the bottleneck for many laboratories and researchers.

Key words: Formicidae, phylogenomics, genotyping-by-sequencing, RADseq, systematics

Inferring the evolutionary relationships among the diversity of organisms on the planet has implications outside of systematics alone and includes fields as diverse as evolutionary biology, ecology, conservation science, food and crop security, and human health. Coupled with this understanding of the importance of having well-resolved evolutionary relationships across the tree of life, new technologies have made possible the generation of data at a scale not feasible in the past. The larger sizes of these new molecular datasets, together with the sequencing of loci scattered across the genome, may make it more likely for researchers to resolve difficult areas of the tree of life. With this comes the opportunity to generate large amounts of phylogenetically informative data for a low cost per nucleotide.

Several reduced-representation genome-sequencing (RRGS) methods are available for phylogenetic inference. These include ultraconserved elements (Faircloth et al. 2012), anchored hybrid enrichment (Lemmon et al. 2012), and several methods that fall under the broad category of restriction-site-associated DNA sequencing (RADseq). These RADseq methods include standard RADseq (RADtag; Baird et al. 2008), double-digest RADseq (ddRAD; Peterson et al. 2012), and genotyping-by-sequencing (GBS; Elshire et al. 2011). One difference between the RADseq methods and GBS is that RADseq requires DNA shearing equipment, which is expensive and not common among smaller molecular laboratories. Another difference is that RADseq methods are likely to sequence almost every restriction site, where GBS results in more missing data across individuals. While outside of the scope of this study to investigate the utility of each of these RRGS methods, we imagine that a direct comparison of these methods for phylogenetic inference will soon be available in the scientific literature, although there are reviews that compare these methods more generally in ecology and evolution (Davey et al. 2011, Andrews et al. 2016). Due to the low cost per sample, lack of special equipment required, and feasibility for most molecular laboratories, we investigate the potential role of GBS in phylogenetic reconstruction for insects. GBS methods result in thousands of 'loci' per sample, although determining homology of these loci between samples can be difficult (Rubin et al. 2012). The length of the loci are often short (50–150 bp), but this is currently limited by the sequencing technologies and can produce longer loci/reads as the technologies improve.

There have been several studies that have implemented GBS at the population and phylogeographic level (Emerson et al. 2010, Elshire et al. 2011, Harvey and Brumfield 2015, Pellegrino et al. 2016); however, fewer empirical studies have leveraged this method at the species level (or higher) (Keller et al. 2013, Escudero et al. 2014, Wong et al. 2015, Winston et al. 2017). One simulated study on real genomic data has demonstrated the potential utility of this and related methods for groups of taxa younger than 50 million years (Mya) (Rubin et al. 2012). To empirically test the utility of the RRGS GBS method for species-rich groups of organisms, we examined two ant genera (*Cephalotes* Latreille 1802 (Hymenoptera: Formicidae) and *Polyrhachis* Smith 1857 (Hymenoptera: Formicidae)) that are distantly related, but have molecular phylogenetic analyses that have been completed in other studies based on modest gene-based approaches. Our goal for this study is to provide an empirical proof-of-concept for the use of RRGS for phylogenetics and outline the protocols we implemented for other laboratories to apply and modify as needed.

Materials and Methods

Taxon Sampling and DNA Extraction

To test the utility of this method, we included 48 samples from our focal taxa groups (Table 1). All samples were collected and stored in ethanol until DNA extractions were performed (Moreau et al. 2013). We selected these groups to represent two distantly related lineages of ants (Moreau et al. 2006, Moreau and Bell 2013), for which previous molecular phylogenies have been inferred using modest gene-based sequencing (Price et al. 2014, Mezger and Moreau 2016, Moreau et al. in prep.) to serve as a point of comparison. In addition, age estimates from previous studies suggest *Cephalotes* and *Polyrhachis* may be appropriate for RADseq phylogenetic inference (25–46 Mya for *Cephalotes*: Price et al. 2014 and Ward et al. 2015; 37–40 Mya for *Polyrhachis*: Blaimer et al. 2015 and Mezger and Moreau 2016). For the analysis of species

Table 1. Details for specimens used in this study

Genus	Species	Specimen accession number	Starting gDNA volume (ng)	Mean depth of coverage	Clusters at 85%	Number of variable loci
Cephalotes	*atratus*	FMNHINS3165140	200	29.6	255,809	20,565
Cephalotes	*atratus*	FMNHINS3165435	200	14.3	181,875	17,665
Cephalotes	*christopherseni*	FMNHINS3165515	200	17.6	107,100	6,030
Cephalotes	*crenaticeps*	FMNHINS3145504	200	10.4	105,595	2,915
Cephalotes	***grandinosus***	FMNHINS3145515	**180**	**7.8**	**23**	**0**
Cephalotes	*maculatus*	FMNHINS3145643	162	15.1	136,733	7,946
Cephalotes	*minutus*	FMNHINS3041069	200	17	253,942	17,092
Cephalotes	***minutus***	FMNHINS3041083	**171**	**7.9**	**14,877**	**205**
Cephalotes	*minutus*	FMNHINS3041093	149	11.6	84,688	2,808
Cephalotes	*minutus*	FMNHINS3145637	200	34.7	362,681	30,187
Cephalotes	***minutus***	FMNHINS3145687	**200**	**8.7**	**67**	**0**
Cephalotes	*mompox*	FMNHINS3145513	200	28.2	196,909	6,968
Cephalotes	*pallens*	FMNHINS3145582	200	11	82,020	3,031
Cephalotes	*pallens*	FMNHINS3145651	200	11.4	83,876	2,300
Cephalotes	***setulifer***	FMNHINS105032	**200**	**9.1**	**27,732**	**424**
Cephalotes	***spinosus***	FMNHINS3165064	**200**	**8**	**15,025**	**212**
Cephalotes	*targionii*	FMNHINS3145513	200	15.6	252,836	18,537
Cephalotes	*unimaculatus*	FMNHINS3145514	200	24.5	294,697	38,541
Cephalotes	*varians*	FMNHINS3041048	122	28.2	343,576	44,086
Cephalotes	*varians*	FMNHINS3144791	200	25	326,667	37,762
Cephalotes	*varians*	FMNHINS3165098	200	14.8	185,619	14,491
Cephalotes	*varians*	FMNHINS3471861	169	13.4	208,489	12,996
Procryptocerus	*balzani*	FMNHINS3165521	200	14.5	225,202	14,825
Procryptocerus	*mayri*	FMNHINS3165522	200	17.7	194,785	10,396
Polyrhachis	*argentosa*	CSM1023	200	20.4	80,456	4,661
Polyrhachis	*brevinoda*	FMNHINS3165516	200	20.3	95,796	5,727
Polyrhachis	*cf elegantula*	CSM1033	200	38.2	104,028	9,502
Polyrhachis	*cf elegantula*	FMNHINS3145194	200	15.4	38,385	1,666
Polyrhachis	*cupreata*	CSM0830	200	24.2	82,187	5,387
Polyrhachis	*cupreata*	CSM1016	200	31.3	96,594	9,616
Polyrhachis	*delecta*	CSM1063	200	34.4	145,429	8,434
Polyrhachis	*delecta*	CSM1078	200	18.3	74,432	3,633
Polyrhachis	*esarata*	FMNHINS3165517	200	42.3	161,653	13,187
Polyrhachis	*foreli*	FMNHINS3165518	200	16.7	58,267	3,058
Polyrhachis	*foreli*	FMNHINS3041069	200	19.3	70,678	3,105
Polyrhachis	*militaris*	FMNHINS105032	200	45.6	138,094	10,746
Polyrhachis	*monteithi*	FMNHINS3145582	200	39.9	123,912	9,203
Polyrhachis	*monteithi*	FMNHINS3165519	200	30.2	83,469	6,041
Polyrhachis	*mucronata*	FMNHINS3041093	200	43.7	111,277	4,076
Polyrhachis	*olybria*	FMNHINS3165520	200	39.4	118,577	17,195
Polyrhachis	*ornata*	FMNHINS3145198	200	25.8	82,160	8,263
Polyrhachis	*robsoni*	FMNHINS3145651	200	31.6	101,416	6,882
Polyrhachis	*robsoni*	CSM0712	200	45.3	129,813	8,947
Polyrhachis	*rufifemur*	FMNHINS3145180	200	14.7	76,544	5,191
Polyrhachis	*senilis*	FMNHINS3145186	200	18	69,639	5,528
Polyrhachis	*senilis*	FMNHINS3145201	200	35.4	156,098	9,380
Polyrhachis	*sokolova*	FMNHINS3145190	200	24.5	166,622	4,798
Camponotus	*novaehollandiae*	FMNHINS3165514	200	35.4	89,818	8,778

Sample details in bold were omitted from our analyses due to low depth of coverage (<10×). Clusters at 85% are the total number of loci after trimming, filtering, and clustering the reads at 85% similarity.

within the turtle ant genus *Cephalotes* (subfamily Myrmicinae), we sampled 22 individuals from 13 different species representing 9 of the 18 'species groups' recognized from Price et al. (2014) plus 2 species of *Procryptocerus*, the sister lineage of *Cephalotes*, to serve as outgroups. For the spiny ants in the genus *Polyrhachis* (subfamily Formicinae), 23 samples were analyzed including 16 species representing 8 of 13 subgenera as well as one species of *Camponotus*, one of the closest known outgroups for *Polyrhachis*. For DNA extractions we followed the Qiagen DNeasy extraction protocol of Moreau (2014) (Qiagen Inc., Valencia, CA). Vouchers for all samples sequenced in this study have been deposited in the scientific collections of the Field Museum of Natural History (Chicago, IL; see Table 1 for voucher codes).

Restriction Enzyme Selection

To determine the most appropriate restriction enzyme, we borrowed from the analysis of Rubin and Moreau (2016), which aligned published ant genomes to identify appropriate restriction sites that in this case were used to provide a linkage map for genome annotation of a single ant species. In previous study, the frequency of cut sites and fragment size for 280 commonly used restriction enzymes (New England BioLabs Inc., Ipswich, MA) was investigated against seven ant genomes (Bonasio et al. 2010; Nygaard et al. 2011; Smith et al. 2011a,b; Suen et al. 2011; Wurm et al. 2011). For this approach, each potential restriction enzyme cut site was blasted against each of the seven ant genomes. To facilitate sequencing of the target fragments, Rubin and Moreau (2016) considered only restriction enzymes that produced a mean distance between cut sites of approximately 1,000 bp. Restriction enzymes that fit these criteria were then compared to previously published studies that have successfully used and created barcode adapters for these restriction enzymes. Based on this approach, we used the restriction enzyme ApeKI (Elshire et al. 2011), but researchers working on other taxonomic groups should preform this step to find the most appropriate restriction enzymes when closely related genomes are available.

GBS Library Preparation

All samples were processed using a modified version of the reduced-representation genome RADseq protocol, GBS, of Elshire et al. (2011) with the addition of a size-selection step (described below) to ensure that all fragments were of an appropriate size to be sequenced on a next-generation sequencing platform.

The restriction enzyme (ApeKI) and adapter and barcoding sequences were all taken from Elshire et al. (2011). We began by quantifying all of our genomic DNA (gDNA) extractions using a High-Sensitivity DNA assay on a Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA), using 2 µl of gDNA per assay. Based on those readings, we calculated the volume required to obtain 200 ng of gDNA for each sample, sometimes using the entire sample if there was not a full 200 ng available (Table 1). The barcode and common adapters were diluted, annealed, quantified, and dried down as in Elshire et al. (2011). We then added 200 ng (if available) or the entire extraction of gDNA to the wells containing the prepared adapters and dried everything down again. These combined and dehydrated samples were then digested with the ApeKI restriction enzyme in 20 µl reactions and adapters were ligated to sticky ends in 50 µl reactions, as in Elshire et al. (2011). Following this protocol, we pooled the barcoded samples by combining 20 µl of each reaction into a single tube and then purified the combined sample using the QIAquick PCR purification kit including the optional sodium acetate (Qiagen Inc).

The pooled libraries were then size selected to a range of 300–800 bp on a 1% high-melt agarose gel and cleaned up with the Qiagen gel extraction kit with a final elution in 30 µl elution buffer (EB) (Qiagen Inc.). Out of concern for overloading the spin columns with gel, we cleaned each size-selected gel slice in its own column and as a result ended up with size-selected samples that were of lower concentrations than we would have generated if we had run all slices through a shared spin column. We compensated for this lower concentration by running more PCRs, as described below. The size-selected library was amplified using the PCR modified from Elshire et al. (2011) with the following thermalcycler profile: 72°C for 5 min, denaturing at 95°C for 1 min, and then 17 cycles of 95°C for 30 s, 65°C for 30 s, and 72°C for 1 min, with a final 5 min elongation step at 72°C. In an attempt to reduce PCR bias, we ran this PCR multiple times as in Rubin and Moreau (2016); however, while that study ran the PCR 4 times, we ran ours 16 times in order to compensate for the low concentration of our size-selected library, although this is not necessary if concentrations are not an issue. PCR products were pooled, and the combined PCR product was then cleaned using a QIAquick PCR purification kit (Qiagen Inc.), with a final elution in 30 µl of EB. The combined sample was pooled with other GBS libraries from our laboratory and sequenced directly on a single lane of an Illumina HiSeq2000 with 100-bp single-end reads. In total, the sample that was sequenced on the HiSeq contained GBS libraries from 192 different individuals to reduce cost per sample while still generating significant data for all samples included.

NGS Data Processing

The resulting next generation sequencing (NGS) Illumina data were analyzed using the pyRAD pipeline, version 2.11 (Eaton and Ree 2013). The analysis was run on a 64-bit Linux workstation with 250 GB of memory and 40 CPUs running at 3 GHz each, though we used only 10 CPUs in each of our analyses. We set the expected sequence of the restriction recognition site overhang to CWGC for ApeKI. Reads were clustered at 85% similarity (both within sample and between samples), and we excluded any sequences that were shorter than 70 bp. Below is a brief description of the different analysis steps carried out by pyRAD, along with the parameters that we specified in our analysis. For a more detailed description, see Eaton (2014).

The first step in pyRAD is demultiplexing of the sequences (i.e., where the raw FASTQ sequences are segregated by their barcode sequences). Step two is a quality-filtering step, in which all barcode and Illumina adapter sequences are removed, and any base calls with a Phred quality score below 33 (the default setting) are changed to *N*. We also set the analyses so that any read with more than four *N*'s was not included in further analyses. In step three, the within-sample sequences are clustered using USEARCH (Edgar 2010), and then the clusters are aligned with MUSCLE (Edgar 2004). Step four uses a maximum-likelihood (ML) equation to estimate both the mean heterozygosity rate and the sequencing error rate from the base counts at each site across all clusters.

The estimates of the mean error rate and heterozygosity from step four are used to create consensus sequences for each cluster in step five. In the process of generating consensus sequences, pyRAD refers to parameters set by the user to filter out sequences based on the following, with the values we used set in parentheses: minimum coverage (6×), maximum number of undetermined sites (four), maximum number of heterozygous sites (five), and maximum number of alleles (two). USEARCH is used again in step six to cluster consensus sequences across samples by sequence similarity. Finally, in step seven, the clustered consensus sequences are aligned, presumed paralogs are removed, and human-readable output is generated in a variety of formats.

Phylogenetic Inference and Assessing Phylogenetic Accuracy

To reconstruct phylogenetic relationships for the two included ant genera, we used the noninterleaved PHYLIP file (Felsenstein 1989) produced by pyRAD. This file contained all of the GBS loci concatenated into a supermatrix, with missing data for any individual sample with incomplete taxon sampling filled in with *N*'s (de Queiroz and Gatesy 2007). We implemented RAxML v.8.1.2 (Stamatakis 2006) using the GTR + GAMMA model of molecular evolution to reconstruct the ML topologies for each of the two datasets. Bootstrap values were estimated from 100 pseudoreplicates starting from random seeds.

The phylogenies that we inferred from our GBS datasets were then compared to respective reference phylogenies based on gene-based Sanger sequencing. The reference phylogeny for *Cephalotes*, from Price et al. (2014), is a Bayesian consensus tree based (although the ML topology of Price et al. (2014) was identical for the taxa included here) on both the combined morphological and molecular dataset and the molecular dataset. The molecular dataset is composed of traditional Sanger sequences of three mitochondrial genes and three protein-coding nuclear genes, for a total concatenated length of 3,479 bp, with 1.2% of the nuclear sequences and 6.9% of the mitochondrial sequences either missing or partially missing. The reference phylogeny for *Polyrhachis*, from Mezger and Moreau (2016), is based on a Bayesian consensus tree (although the ML topology was identical for the taxa included here) generated from traditional Sanger sequence dataset consisting of two mitochondrial genes and six protein-coding nuclear genes. The total concatenated length of this dataset was 4,923 bp, with 6.4% of the nuclear sequences and 16.6% of the mitochondrial sequences either missing or partially missing. These reference phylogenies were pruned using Mesquite (Maddison and Maddison 2016) to match the same taxa and number of included samples.

Results

GBS Library Preparation and NGS Data Processing

After implementing the RRGS GBS sequencing protocol of Elshire et al. (2011) with the addition of a size selection step our Illumina HiSeq run yielded an average of 4.03×10^6 reads for the 19 *Cephalotes* and *Procryptocerus* samples (5 samples out of the original 24 failed to produce a mean depth of coverage of at least 10× and so were omitted from analysis; see Table 1), which resulted in an average of 2.04×10^5 loci per sample after filtering and clustering with a minimum of 6× coverage. The mean depth of coverage for the *Cephalotes* dataset was 18.6. The minimum taxa dataset (i.e., the set of loci for which there were at least four samples with representative data) for this group had an average of 2.5×10^4 consensus loci. The average ML estimate of the sequencing error rate was 1.81×10^{-3} and the average heterozygosity rate was 8.03×10^{-3}. In our final matrix, the total number of variable loci was 309,141 and of those there were 54,214 that were parsimony informative.

For the analysis of the GBS libraries prepared from the *Polyrhachis* and *Camponotus* samples, an average of 2.88×10^6 reads were sequenced, generating on average 1.02×10^5 loci per sample after filtering and clustering. The mean depth of coverage for these samples was 29.6, and there were an average of 3.4×10^4 loci in each sample's minimum taxa dataset. The average ML estimate of the sequencing error rate for these samples was 1.58×10^{-3}, and the average rate of heterozygosity was 3.92×10^{-3}. Our final dataset had 173,004 variable loci, of which 60,279 are parsimony informative. See Table 2 for a summary of these data and all data have been deposited in NCBI (BioProject ID: PRJNA383731).

Accounting for the costs of DNA extraction, GBS library preparation, and multiplexing our 48 samples in one Illumina HiSeq lane (along with three other GBS libraries from our laboratory for a total of 192 samples run on a single HiSeq2000 lane), our cost per sample was approximately US$12 (Table 3). This cost does not account for any labor costs associated with the library preparation as we prepared the libraries in our laboratory. In addition, there is a large up-front cost for the barcoded adapters (~US$1,100.00), but the volume purchased is enough for thousands of reactions per barcode. After the cost of sequencing on the HiSeq2000, the largest cost for this method was the DNA extraction with the Qiagen blood and tissue kit. Given the sensitivity of restriction enzymes to contaminants sometimes found in DNA extractions, we do not recommend lower cost extraction methods when possible.

Phylogenetic Inference

Cephalotes

Our analyses recovered as monophyletic all *Cephalotes* turtle ant species for which we had multiple samples represented, and the two *Procryptocerus* cluster together outside of the *Cephalotes* clade. *Cephalotes* lineages had been previously grouped into 18 different 'species groups' based on morphological data (de Andrade and Baroni Urbani 1999). In all instances in this study where there were multiple representative species from the same species group, our analysis recovered the group as being monophyletic. This topology matches that of the phylogeny from Price et al. (2014), which was based on Sanger gene-based sequencing combined with morphological data.

Table 2. Summary of sequencing results by taxonomic group

Taxonomic group	Mean number loci	Mean depth of coverage per locus	Mean number variable loci per sample	Minimum taxa dataset mean size	Average Heterozygosity	Average number of parsimony-informative sites across aligned matrix
Cephalotes/*Procryptocerus* (*n* = 19)	204,373	19	16,271	25,273	0.00803	54,214
Polyrhachis/*Camponotus* (*n* = 24)	102,306	30	7,209	34,530	0.00392	60,279

Taxonomic group is the target group plus the outgroup. Five samples were excluded from the original 24 *Caphalotes*/*Procryptocerus* samples due to low depth of coverage (see Table 1). Mean number of loci passing quality filter is the average number of clusters formed at 85% similarity, after sequences were trimmed and filtered. The minimum taxa dataset of a sample is the set of loci for which there were at least four samples with representative data. Average number of parsimony-informative variable sites is the total number for each respective taxonomic group.

Results of our analyses also lend strong support (bootstrap support > 0.95) to some nodes, where support was lacking (posterior probability < 0.95) in the Sanger/morphology tree for both the Bayesian and ML topologies (Figs. 1 and 2). All nodes in our GBS tree were well supported, while there were three nodes in the trimmed Sanger/morphology tree of Price et al. (2014) for which support was lacking (Fig. 2). For example, the node representing the split of the *basalis* group from the clade including the *laminatus/pusillus*, *pallens*, *pinelli*, *coffee,* and *angustus* groups lacks strong support in the Sanger/morphology topology but has high support in our GBS phylogeny (Fig. 1).

However, there were several differences in well-supported relationships between the species groups in the Sanger/morphology tree of Price et al. (2014) and our GBS topology. For example, in

Table 3. Cost per sample for GBS prep and sequencing

Reagent (or kit)	Company	Total cost	Amount purchased	Amount used per sample	Cost per sample
Extraction Kit	Qiagen	$677	250	1 sample	$2.71
ApeKI RE	New England Biolabs	$266	500 units (250 µl)	1 µl	$1.06
QIAquick PCR Purification Kit	Qiagen	$524	250	Used two for the whole library prep	$0.09
Gel Extraction Kit	Qiagen	$527	250	1 sample for whole library	$0.04
HiSeq Run	Illumina	$1,088	1 lane	1/192 lane	$5.67
NEB Buffer 3 (1x)	New England Biolabs	$19	5 mL		
T4 ligase	New England Biolabs	$64	20,000 units	640 units	$2.05
Taq Master Mix	New England Biolabs	$140	500 reactions	1 reaction	$0.28
Primers	IDT	$1,100	approx. 300 µl per primer	1 reaction per sample	Negligible
Total cost					$11.90

Prices based from quotes from June 2016.

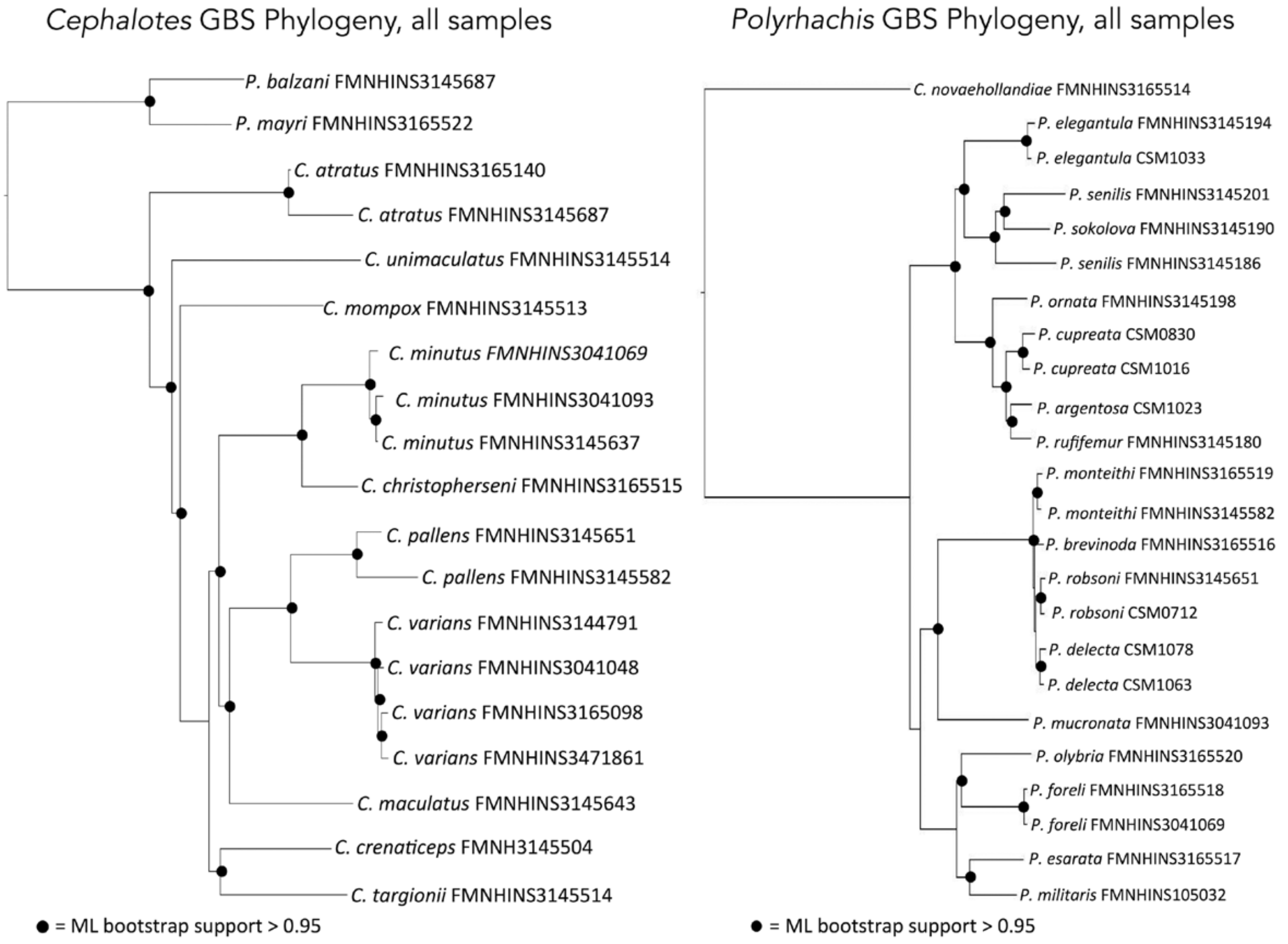
Cephalotes GBS Phylogeny, all samples
P. balzani FMNHINS3145687
P. mayri FMNHINS3165522
C. atratus FMNHINS3165140
C. atratus FMNHINS3145687
C. unimaculatus FMNHINS3145514
C. mompox FMNHINS3145513
C. minutus FMNHINS3041069
C. minutus FMNHINS3041093
C. minutus FMNHINS3145637
C. christopherseni FMNHINS3165515
C. pallens FMNHINS3145651
C. pallens FMNHINS3145582
C. varians FMNHINS3144791
C. varians FMNHINS3041048
C. varians FMNHINS3165098
C. varians FMNHINS3471861
C. maculatus FMNHINS3145643
C. crenaticeps FMNH3145504
C. targionii FMNHINS3145514
● = ML bootstrap support > 0.95
Polyrhachis GBS Phylogeny, all samples
C. novaehollandiae FMNHINS3165514
P. elegantula FMNHINS3145194
P. elegantula CSM1033
P. senilis FMNHINS3145201
P. sokolova FMNHINS3145190
P. senilis FMNHINS3145186
P. ornata FMNHINS3145198
P. cupreata CSM0830
P. cupreata CSM1016
P. argentosa CSM1023
P. rufifemur FMNHINS3145180
P. monteithi FMNHINS3165519
P. monteithi FMNHINS3145582
P. brevinoda FMNHINS3165516
P. robsoni FMNHINS3145651
P. robsoni CSM0712
P. delecta CSM1078
P. delecta CSM1063
P. mucronata FMNHINS3041093
P. olybria FMNHINS3165520
P. foreli FMNHINS3165518
P. foreli FMNHINS3041069
P. esarata FMNHINS3165517
P. militaris FMNHINS105032
● = ML bootstrap support > 0.95

Fig. 1. ML molecular phylogenies generated for this study using the RRGS method GBS. On the left is a phylogeny of the ant genus *Cephalotes* and the outgroup *Procryptocerus.* This *Cephalotes* topology including all samples that had sufficient data generated for our analyses. On the right is the full phylogeny of all samples included for our analysis of the ant genus *Polyrhachis,* along with the outgroup *Camponotus.* Both phylogenies were inferred from Illumina data generated from GBS libraries and were inferred using RAxML. Nodes with ML bootstrap support >0.95 are represented by black dots on nodes.

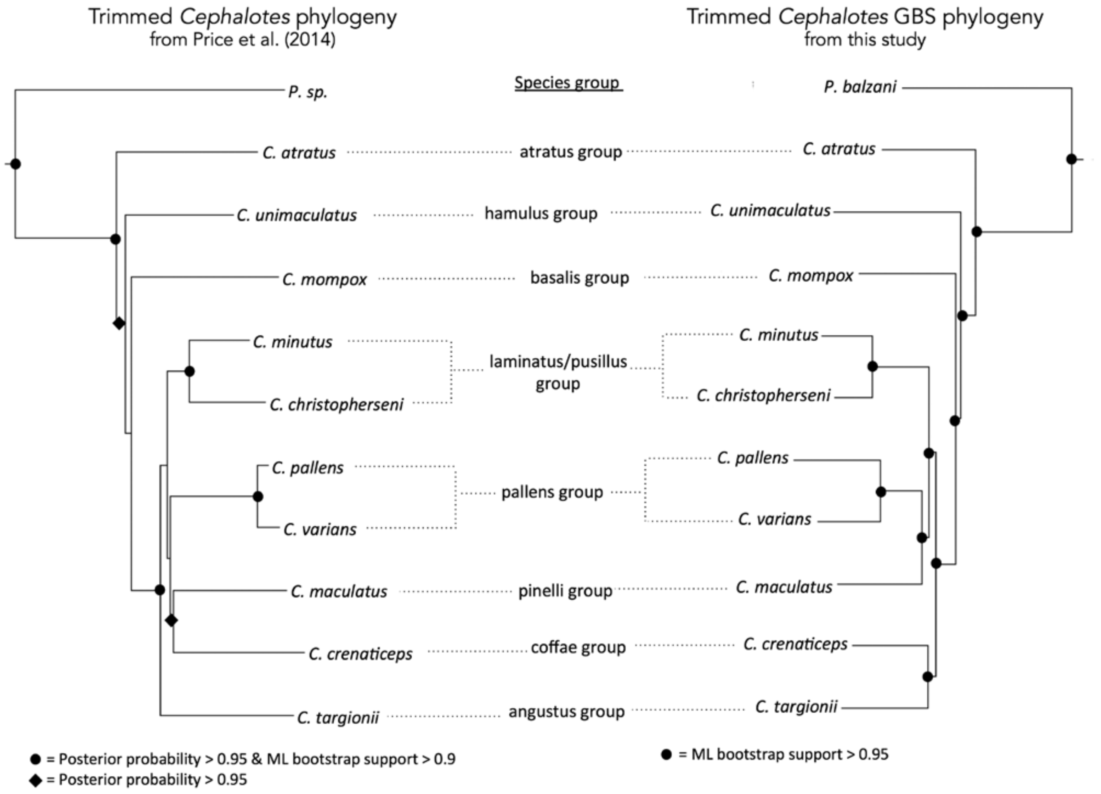

Fig. 2. Comparison between phylogenetic topologies generated by different molecular methods for the turtle ant genus *Cephalotes.* On the left is the tree inferred by Price et al. (2014), trimmed with the Mesquite prune clade tool to represent only the taxa represented in the present study. The Price et al. (2014) phylogeny was inferred with Sanger DNA sequence data from three protein-coding nuclear genes (1,457 bp) and three mitochondrial genes (2,022 bp). On the right is the phylogeny inferred for this study, also pruned with the Mesquite prune clade tool to remove duplicate taxa. This GBS topology was inferred from a dataset that had on average 25,000 loci per sample, each 100 bp in length. In the middle are the morphological species groups assigned by de Andrade and Baroni Urbani (1999). Nodes with Bayesian posterior probabilities >0.95 and/or ML bootstrap support >0.9 are represented by black dots on nodes; Nodes with only Bayesian posterior probabilities >0.95 and ML bootstrap support >0.7 are represented with black diamonds on nodes.

the phylogeny of Price et al. (2014), *Cephalotes targionii* (*angustus* group) is sister to the clade containing the *laminatus*/*pusillus*, *pallens*, *pinelli*, and *coffae* groups, with the latter two groups sister to each other. However, in the GBS phylogeny from this study, *C. targionii* is sister to *Cephalotes crenaticeps* (*coffae* group, bootstrap support > 0.95), and the two together are sister to a clade containing the *laminatus*/*pusillus*, *pallens*, and *pinelli* groups in a strongly supported split (Figs. 1 and 2).

Polyrhachis

All four subgenera for which we had multiple representative species were recovered as monophyletic. Recent work by Mezger and Moreau (2016) challenged the monophyly of several of the previously defined subgenera and reconfigured the genus-wide topology into four broad regions composed of multiple subgenera, numbered I through IV, based on the Sanger-based molecular phylogeny they generated. Our study did not include multiple species of any of the subgenera challenged in the Mezger and Moreau (2016) study; however, we were able to recover the three geographic regions of the tree included in this study as being monophyletic, as found in the previous study.

The topology of the GBS phylogeny generated for this study was congruent with the phylogeny from Mezger and Moreau (2016) with the exception of the relationships within the subgenus *Cyrtomyrma* (within region II). In both trees, the basal node of the group is highly supported, but the splits differ between the two trees. In the phylogeny of Mezger and Moreau (2016), the *Cyrtomyrma* subgenus is inferred with *Polyrhachis delecta* as sister to the rest of the clade, however in our GBS topology the split is between *Polyrhachis monteithi* and the rest of the subgenus. There are also differences in other nodes, though none of those nodes are supported in either topology. For all species where we had multiple representatives, we recovered them as monophyletic with the exception of *Polyrhachis senilis* with *Polyrhachis sokolova* nested within this clade (Figs. 1 and 3).

Discussion

The development of high-throughput sequencing technologies has removed data acquisition as the primary bottleneck for empirical phylogeneticists. As DNA sequencing technologies continue to make sequencing vast amounts of data more affordable, even for non-model organisms, ensuring that these data are appropriate for the level of question being addressed is critical. To address the utility of RRGS for phylogenetic inference, we used the RADseq GBS method of Elshire et al. (2011) with the addition of a size selection step for two distantly related ant genera of varying ages. Our results demonstrate not only the effectiveness of this method for species level phylogenetics, at least for the two ant genera we included but also the cost effectiveness and technical feasibility of such an undertaking for smaller research laboratories.

On average we recovered 150,000 loci and an average of 57,000 parsimony-informative sites per taxonomic group, demonstrating the large amount of data that are available for phylogenetic inference using the GBS RADseq method we implemented

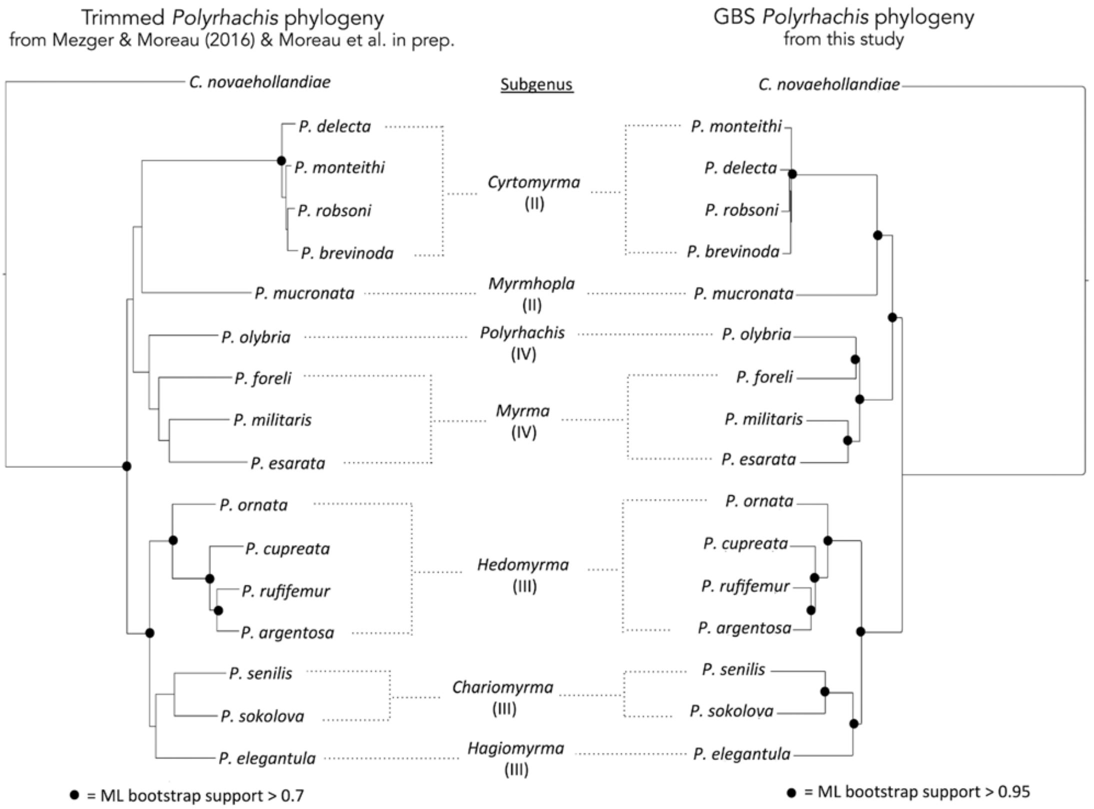

Fig. 3. Comparison between phylogenetic topologies generated by different molecular methods for the spiny ant genus *Polyrhachis*. On the left is the tree inferred by Moreau et al. (in prep.), trimmed with the Mesquite prune clade tool to represent only the taxa represented in the present study. The Moreau et al. (in prep.) topology was inferred with Sanger DNA sequence data from two mitochondrial genes (1,995 bp) and six protein-coding nuclear genes (2,928 bp). On the right is the phylogeny inferred for this study, also pruned with the Mesquite prune clade tool to remove duplicate taxa. This GBS topology was inferred from a dataset that had on average 34,000 loci per sample, each 100 bp in length. In the middle are the subgenera, with taxonomic groups assigned from Mezger and Moreau (2016) in parentheses. Nodes in the Sanger phylogeny with ML bootstrap support >0.7 and nodes in the GBS phylogeny with ML bootstrap support >0.95 are represented by black dots on nodes.

here. Our cost per sample was $12, most of which was incurred by the sequencing step, which could be further reduced with additional barcode adapters for additional multiplexing. We believe this method will be a viable option of many researchers interested in species, or higher level, phylogenetic relationships of non-model organisms, especially if sequencing costs continue to drop near the rate at which they have dropped over the last 10 years (Wetterstrand 2015).

In our analyses of the phylogenetic relationships of two distantly related ant genera, *Cephalotes* and *Polyrhachis*, we recovered highly resolved topologies. For both genera, we were able to resolve nodes that had previously been difficult to resolve and/or recover statistical support with previous gene-based Sanger sequencing datasets (Figs. 2 and 3). For the *Cephalotes* phylogeny, we also recovered maximum clade support for all nodes, while there were two nodes that did not receive statistical support in the *Polyrhachis* phylogeny. There has been work (Rubin et al. 2012) suggesting that this type of reduced-representation genomic sequencing data may be insufficient for resolving clades representing splits that occurred more than 50 Mya because of a drop out of loci due to mutation at restriction sites or difficultly assigning homology between variable loci. Although evolutionary age does not seem to be the cause, we cannot rule this out for the parts of our topology that lack support, but regardless, overall the topologies between our RRGS data and the previous studies are largely congruent and in some cases our data resolves parts of the topologies that lacked support in the gene-based sequencing studies.

One limitation of this method is that as a restriction-based method, data from one dataset are not likely to be easily combinable with other studies due to issues with homology. For this reason, more expensive methods based on target-capture-based RRGS may be preferred. In addition, we caution using this method on taxonomic groups that are older than 50 Mya as this may affect the reliability of the data to infer phylogenetic relationships. Finally, efforts previously directed at generating molecular data should now be shifted toward developing bioinformatics tools and computational biology skills for the analysis of the large volume of data currently being produced.

Acknowledgments

We thank Benjamin E.R. Rubin for helpful advice in implementing all aspects of this project. We also thank Max E. Winston and Felix Grewe for helpful conversations and advice with data analysis and Paul Grabowski for initial assistance with the GBS molecular laboratory protocol. We thank Shauna L. Price for sharing her *Cephalotes* data and results. We also thank two anonymous reviewers and Editor, Sydney Cameron, for very helpful suggestions to improve this manuscript. The molecular portion of this research was completed in the Pritzker Laboratory for Molecular Systematics and Evolution at the Field Museum of Natural History, Chicago, IL, and supported in part by the Negaunee Foundation and the National Science Foundation (NSF DEB 1442316) to C.S.M.

References Cited

de Andrade M. and C. Baroni Urbani. 1999. Diversity and adaptation in the ant genus Cephalotes, past and present. Stuttgarter Beiträge zur Naturkunde Serie B (Geologie und Paläontologie) 271: 1–889.

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

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376.

Blaimer, B. B., S. G. Brady, T. R. Schultz, M. W. Lloyd, B. L. Fisher, and P. S. Ward. 2015. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. BMC Evol. Biol. 15: e271.

Bonasio, R., G. Zhang, C. Ye, N. S. Mutti, X. Fang, N. Qin, G. Donahue, P. Yang, Q. Li, C. Li, et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator.* Science 329: 1068–1071.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12: 499–510.

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

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

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

Edgar R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461.

Emerson, K. J., C. R. Merz, J. M. Catchen, P. A. Hohenlohe, W. A. Cresko, W. E. Bradshaw, and C. M. Holzapfel. 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl. Acad. Sci. USA 107(37): 16196–16200.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. PLoS One, 6: e19379.

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

Faircloth, B. C., J. E. McCormack, N. G. Crawford, et al. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61: 717–726.

Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.

Harvey, M. G. and R. T. Brumfield. 2015. Genomic variation in a widespread Neotropical bird (*Xenops minutus*) reveals divergence, population expansion, and gene flow. Mole. Phylogenet. Evol. 83: 305–316.

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

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61: 727–744.

Maddison, W. P. and D.R. Maddison. 2016. Mesquite: a modular system for evolutionary analysis. Version 3.11 http://mesquiteproject.org

Mezger, D. and C. S. Moreau. 2016. Out of South-East Asia: phylogeny and biogeography of the spiny ant genus *Polyrhachis* (Hymenoptera: Formicidae). Syst. Entomol. 41: 369–378.

Moreau, C. S. 2014. A practical guide to DNA extraction, PCR, and gene-based DNA sequencing in insects. Halteres 5: 32–42.

Moreau, C. S. and C. D. Bell. 2013. Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. Evolution 67: 2240–2257.

Moreau, C. S., C. D. Bell, R. Vila, S. B. Archibald, and N. E. Pierce. 2006. Phylogeny of the ants: diversification in the age of angiosperms. Science 312: 101–104.

Moreau, C. S., B. D. Wray, J. E. Czekanski-Moir, and B. E. R. Rubin. 2013. DNA preservation: A test of commonly used preservatives for insects. Invert. Syst. 27: 81–86.

Nygaard, S., G. Zhang , M. Schiott, C. Li, Y. Wurm, H. Hu, J. Zhou, L. Ji, F. Qiu, M. Rasmussen. et al. 2011. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. Genome Res. 21: 1339–1348.

Pellegrino, I., L. Boatti, M. Cucco, F. Mignone, T. N. Kristensen, N. Mucci, E. Randi, A. Ruiz-Gonzalez, and C. Pertoldi. 2016. Development of SNP markers for population structure and phylogeography characterization in little owl (*Athene noctua*) using a genotyping- by-sequencing approach. Conserv. Gene. Res. 8: 13.

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

Price, S. L., S. Powell, D. J. C. Kronauer, L. A. P. Tran, N. E. Pierce, and R. K. Wayne. 2014. Renewed diversification is associated with new ecological opportunity in the Neotropical turtle ants. J. Evol. Biol. 27: 242–258.

de Queiroz, A. and J. Gatesy. 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22: 34–41.

Rubin, B. E. R. and C. S. Moreau. 2016. Comparative genomics reveals convergent rates of evolution in ant–plant mutualisms. Nature Comm. 7: 12679.

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

Smith, C. R., C. D. Smith, H. M. Robertson, M. Helmkampf, A. Zimin, M. Yandell, C. Holt, H. Hu, E. Abouheif, R. Benton. et al. 2011a. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc. Natl. Acad. Sci. USA* 108: 5667–5672.

Smith, C. D., A. Zimin, C. Holt, E. Abouheif, R. Benton, E. Cash, V. Croset, C. R. Currie, E. Elhaik, C. G. Elsik, et al. 2011b. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proc. Natl. Acad. Sci. USA* 108: 5673–5678.

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

Suen G., C. Teiling, L. Li, C. Holt, E. Abouheif, E. Bornberg-Bauer, et al. 2011. The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. PLOS Genet. 7: e1002007.

Ward, P. S., S. G. Brady, B. L. Fisher, and T. R. Schultz. 2014. The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae). System. Entomol. 40: 61–81.

Wetterstrand, K. A. 2015. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at www.genome.gov/sequencingcosts. (Accessed 7 January 2015).

Winston, M. E., D. Kronauer, and C. S. Moreau. 2017. Early and dynamic colonization of Central America drives speciation in Neotropical army ants. Mol. Ecol. 26: 859–870.

Wong, M. M. L., N. Gujaria-Verma, L. Ramsay, H. Y. Yuan, C. Caron, M. Diapari, et al. 2015. Classification and characterization of species within the genus Lens using Genotyping-by-Sequencing (GBS). PLoS One 10: e0122025.

Wurm, Y., J. Wang, O. Riba-Grognuz, M. Corona, S. Nygaard, B. G. Hunt, K. K. Ingram, L. Falquet, M. Nipitwattanaphon, D. Gotzek, et al. 2011. The genome of the fire ant *Solenopsis invicta*. Proc. Natl. Acad. Sci. USA 108: 5679–5684.