Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017 doi: 10.1111/2041-210X.12742

Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera

Michael G. Branstetter*,[1,2], John T. Longino[1], Philip S. Ward[3] and Brant C. Faircloth*,[4]

[1]Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112, USA; [2]Department of Entomology, National Museum of Natural History, Smithsonian Institution, PO Box 37012, 10th & Constitution Aves. NW, Washington, D.C. 20560, USA; [3]Department of Entomology and Nematology, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA; and [4]Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA 70803, USA

Summary

1. Targeted enrichment of conserved genomic regions (e.g. ultraconserved elements or UCEs) has emerged as a promising tool for inferring evolutionary history in many organismal groups. Because the UCE approach is still relatively new, much remains to be learned about how best to identify UCE loci and design baits to enrich them.

2. We test an updated UCE identification and bait design workflow for the insect order Hymenoptera, with a particular focus on ants. The new strategy augments a previous bait design for Hymenoptera by (i) changing the parameters by which conserved genomic regions are identified and retained, and (ii) increasing the number of genomes used for locus identification and bait design. We perform *in vitro* validation of the approach in ants by synthesizing an ant-specific bait set that targets UCE loci and a set of ‘legacy’ phylogenetic markers. Using this bait set, we generate new data for 84 taxa (16/17 ant subfamilies) and extract loci from an additional 17 genome-enabled taxa. We then use these data to examine UCE capture success and phylogenetic performance across ants. We also test the workability of extracting legacy markers from enriched samples and combining the data with published datasets.

3. The updated bait design (hym-v2) contained a total of 2590-targeted UCE loci for Hymenoptera, significantly increasing the number of loci relative to the original bait set (hym-v1; 1510 loci). Across 38 genome-enabled Hymenoptera and 84 enriched samples, experiments demonstrated a high and unbiased capture success rate, with the mean locus enrichment rate being 2214 loci per sample. Phylogenomic analyses of ants produced a robust tree that included strong support for previously uncertain relationships. Complementing the UCE results, we successfully enriched legacy markers, combined the data with published Sanger datasets and generated a comprehensive ant phylogeny containing 1060 terminals.

4. Overall, the new UCE bait design strategy resulted in an enhanced bait set for genome-scale phylogenetics in ants and other Hymenoptera. Our *in vitro* tests demonstrate the utility of the updated design workflow, providing evidence that this approach could be applied to any organismal group with available genomic information.

Key-words: Formicidae, molecular systematics, next-generation sequencing, phylogenomics, targeted enrichment, ultraconserved elements

Introduction

Advances in sequencing technology and laboratory protocols have invigorated phylogenetic systematics (Lemmon & Lemmon 2013; McCormack *et al.* 2013b; Yeates *et al.* 2016). By coupling next-generation sequencing with genomic reduction and sample multiplexing, it has become increasingly feasible to generate genome-scale datasets from hundreds of samples within a short timeframe. Among competing approaches (e.g. transcriptomics, RADseq and targeted enrichment), the targeted enrichment of conserved or ultraconserved elements (i.e. UCEs, *sensu* Faircloth *et al.* 2012) has grown rapidly in popularity, with the technique being applied to a variety of organisms, including birds (McCormack *et al.* 2013a), mammals (McCormack *et al.* 2012), fish (Faircloth *et al.* 2013), reptiles (Crawford *et al.* 2012) and, most recently, arthropods (Blaimer *et al.* 2015, 2016a; Faircloth *et al.* 2015; Branstetter *et al.* 2016a,b; Starrett *et al.* 2016). The reason for this rapid growth is due to the many positive attributes of the UCE approach: it works with suboptimally preserved specimens and/or degraded DNA (McCormack, Tsai & Faircloth 2015; Blaimer *et al.*

*Correspondence authors. E-mail: mgbranstetter@gmail.com; brant@faircloth-lab.org

2016b), it results in data that can resolve ancient (Crawford *et al.* 2012) and recent divergence events (Harvey *et al.* 2016; Manthey *et al.* 2016), and it costs relatively little (in time and money) for the amount of data generated. The UCE approach also benefits from a user group that freely shares bait sets, lab protocols (see www.ultraconserved.org) and bioinformatics tools (Faircloth 2016), making it an easy method to learn and use, compared to similar approaches.

Despite the overall success of the UCE approach (*sensu* Faircloth *et al.* 2012), uncertainty exists regarding how to improve and optimize various steps of the process (Glenn & Faircloth 2016), including UCE identification and bait design. Thus, testing new UCE design strategies is of broad interest and is important for refining the technique. Here, we introduce an improved bait set for the insect order Hymenoptera, a megadiverse lineage that includes sawflies, parasitoid wasps, stinging wasps, bees and ants. The main objectives of the study are to further test a new workflow for UCE marker identification and bait design (detailed in Faircloth in press) and ultimately to enhance a recently published bait set for Hymenoptera (Faircloth *et al.* 2015), with particular emphasis on improving capture success in ants. In the design of the first Hymenoptera bait set ('hym-v1'), Faircloth *et al.* (2015) employed the standard UCE bait design workflow (detailed in Faircloth *et al.* 2012). Using this approach, the authors identified 1510 UCE loci by aligning and scanning the genomes of two species, the parasitoid wasp *Nasonia vitripennis* and the honeybee *Apis mellifera*. To enrich these loci in non-model taxa, the authors designed 2749 baits from a single hymenopteran genome (*N. vitripennis*). The bait set was tested and validated using both *in silico* and *in vitro* analyses.

Although the hym-v1 bait set has been successfully used in multiple studies (Blaimer *et al.* 2015, 2016a; Faircloth *et al.* 2015; Branstetter *et al.* 2016a,b), the bait design could be modified to improve data quality. One limitation of hym-v1 is that the capture success rate is highly variable among lineages, with success decreasing at increasing phylogenetic distances from *Nasonia* (Faircloth *et al.* 2015; Branstetter *et al.* 2016b). Enrichment success is particularly poor among sawfly lineages; likely because no sawfly genomes were available when the bait set was designed. Another potential shortfall of hym-v1 is that it has been necessary to allow for modest amounts of missing data to retain a large number of loci (taxon occupancy as low as 75%). While this is not an alarming amount of missing data by current phylogenomic standards, reducing missing data from phylogenetic analyses is generally desirable. We try to address all these limitations for Hymenoptera by employing a new UCE bait design strategy that includes the following improvements: (i) we change the parameters of how UCEs are identified and retained to increase the number of loci (i.e. we employ the updated bait design workflow of Faircloth in press), and (ii) we increase the number of hymenopteran genomes used for both UCE marker identification and bait design. Designing the baits from multiple hymenopteran genomes increases bait complexity, which should increase enrichment success, and allows researchers to synthesize either the entire bait set or taxon-specific bait subsets.

To validate our new design approach *in vitro*, we focus on one group within Hymenoptera, the ants (Formicidae), a lineage that originated in the Cretaceous and diversified into more than 14 000 extant species (Ward 2014). We started by synthesizing a new, ant-specific bait set, in which we included baits designed from two distantly related ant genomes. To make this updated bait set 'back compatible' with data from the original bait set and with data from 'legacy' phylogenetic markers, we included new baits targeting loci present in the original (hym-v1) bait set (Faircloth *et al.* 2015) as well as baits targeting 16 nuclear genes that have been commonly sequenced in insects. Following bait synthesis, we performed *in vitro* tests of the new bait set on all but one ant subfamily and several outgroups. We then used the sequence data to evaluate (i) capture success of the new bait set, and (ii) phylogenetic performance in ants. We also (iii) demonstrate how the inclusion of exon baits made it possible to combine sequence data from enriched samples with published datasets based on 'legacy' markers. Overall, our study demonstrates the utility of the updated design workflow (Faircloth in press), providing evidence that this approach could be applied to any organismal group with available genomic information.

Materials and methods

UCE BAIT DESIGN

To identify conserved genomic regions and design enrichment baits targeting those regions we employed an updated UCE bait design workflow (described in detail in Faircloth in press). The main difference between the updated workflow and the standard UCE workflow (*sensu* Faircloth *et al.* 2012) is that instead of performing synteny-based, genome-genome alignment to identify UCEs, short reads from multiple genomes are permissively mapped to a single base genome. The new approach greatly increases the number of identified conserved regions (Faircloth in press).

We simulated random reads from the genomes of six 'exemplar' hymenopterans (Table S1) using ART (Huang *et al.* 2012); for each genome we aligned the reads to a base genome sequence of the sawfly *Athalia rosae* (Hymenoptera: Tenthredinidae) using STAMPY (Lunter & Goodson 2011) with 'substitutionrate=0·05', 'insertsize=400', 'maxbasequal=93' and four compute threads; and we converted alignments to BAM format using SAMTOOLS (Li *et al.* 2009). Following alignment, we used SAMTOOLS to remove unaligned reads from the BAM alignment files, and we used BEDTOOLS (Quinlan & Hall 2010) to convert BAM files to BED format, sort the resulting BED files and merge proximate (<100 bp distance) intervals. Then, we used a program (phyluce_probe_strip_masked_loci_from_set) from the PHYLUCE v1.6 package (Faircloth 2016; all programs beginning with 'phyluce_' in the name are available within the PHYLUCE v1.6 package) to screen the resulting interval data and remove intervals in each BED file that overlapped repeat-masked, short (<80 bp), or ambiguous segments of the *A. rosae* genome. The intervals that were not filtered represent conserved sequences (<5% sequence divergence) shared between the base genome and at least one of the exemplar taxa. We then used another program (phyluce_probe_get_multi_merge_table) to determine which of these conserved intervals were shared among *A. rosae* and all of the exemplar taxa. We buffered these regions to 160 bp and extracted sequence from the *A. rosae* genome corresponding to the buffered intervals. We then designed a file of 120 bp temporary bait sequences from the *A. rosae* genome by tiling two baits over the centre of each sequence (baits overlapped by 80 bp). We aligned the baits to each other using LASTZ (phyluce_probe_easy_lastz), and we screened the file of temporary bait sequences to remove duplicate baits that were >50% identical over >50% of their length (phyluce_probe_remove_duplicate_hits_from_probes_using_lastz).

To design a more diverse final bait set that incorporated bait sequences from the exemplar species, we aligned the file of filtered temporary baits to the genomes of each exemplar taxon plus *A. rosae* using a wrapper (phyluce_probe_run_multiple_lastzs_sqlite) around LASTZ (Harris 2007) with liberal alignment parameters (≥50% sequence identity required to map). To avoid possible paralogs, we removed loci that were hit by baits targeting different conserved regions (loci) or different loci that were hit by the same bait, and we extracted FASTA sequences at each locus (buffered to 160 bp) from all assemblies. We used another program (phyluce_probe_get_multi_fasta_table) to compute which loci we detected across the hymenopteran genome assemblies and created a list of those loci detected in five of the seven assemblies. We then designed a second temporary bait set targeting these loci by tiling baits across each locus in each of the seven hymenoptera genomes where we detected the locus. As above, we aligned the baits to each other and screened the resulting bait set to remove duplicate baits that were >50% identical over >50% of their length.

We combined these new baits with an earlier bait set (hym-v1; Faircloth *et al.* 2015) by removing those version 1 loci that performed poorly (captured in <75% of taxa tried; data from Branstetter *et al.* 2016b) and removing those version 2 baits that were already present in the version 1 set. Then, based on bait positions in the *N. vitripennis* and *A. mellifera* genomes, we removed baits targeting loci that were <1 kbp from each other. We then re-designed the set of version 1 baits following steps similar to those detailed above by aligning the reduced version 1 baits to the exemplar genomes, removing possible duplicate hits, extracting FASTA sequences at each locus, determining which loci we detected in five of the seven assemblies, and tiling baits across each locus in each of the seven hymenoptera genomes where we detected the locus. Finally, we combined these updated version 1 baits with the new baits designed above, and we screened these baits once again for duplicates. We called this bait set, designed to work universally across Hymenoptera, the 'principal hym-v2 bait set'. We used a program in PHYLUCE (phyluce_probe_get_subsets_of_tiled_baits) to select the subset of these hymenopteran baits that were designed from the ant genomes (*Atta cephalotes*, *Harpegnathos saltator*). We called this file the 'ant-specific hym-v2 bait set.'

EXON BAIT DESIGN

To enable the integration of sequence data collected using the new UCE bait set with data generated by targeted PCR and Sanger sequencing, we designed baits to enrich 16 exons from 12 commonly sequenced nuclear genes in ants (Table S2; *AbdA*, *Antp*, *ArgK*, *CAD*, *EF1α-F1*, *EF1α-F2*, *LwRh*, *NaK*, *POLD1*, *Top1*, *Ubx*, *Wg*). We began the design process by aligning exon sequences from four, distantly related ant species (*Martialis heureka*, *Leptogenys diminuta*, *Cerapachys jacobsoni*, *Myrmica tahoensis*) to one another, and for each exon and ant species, we generated 120 bp baits evenly tiled across the aligned exon sequence, resulting in a tiling density of ~2× (baits overlapped by ~60 bp). We designed baits from multiple ant species in order to introduce variability into the individual baits and increase enrichment success of each exon across all ant species. We added the resulting 452 exon baits to the ant-specific hym-v2 bait set, described above, and we had the baits synthesized by MYcroarray (MYcroarray, Ann Arbor, MI, www.mycroarray.com).

TAXON SELECTION FOR ANT PHYLOGENOMICS

We selected a total of 101 taxa for inclusion in our phylogenomic study, generating new sequence data for 84 taxa (Table S3) and extracting data from available genomes for 17 taxa (Table S1). We selected taxa to test the ability of the new bait set to evenly enrich UCE loci across major ant lineages and outgroups. We also wanted to investigate phylogenetic relationships among ant subfamilies and tribes within the subfamily Myrmicinae using phylogenomic data. Among ants, we included representatives of all subfamilies, except for the extremely rare Martialinae. Outside of ants, we included eight species from five groups within the stinging wasps (Aculeata): Vespidae, Pompilidae, Tiphiidae, Sphecidae and Anthophila (bees).

WET LAB PROTOCOL

The protocol for capturing and sequencing UCE loci closely followed the methods reported in Faircloth *et al.* (2015), but is described in detail in Appendix S2.

EXTRACTION OF UCE LOCI FROM GENOME-ENABLED TAXA

To examine the presence of the newly identified UCE loci across a diverse set of hymenopteran species, we used *in silico* methods (see Faircloth 2016) to identify and extract the updated set of UCE loci from 38 hymenopteran genomes, including 13 ant genomes (Table S1). We obtained each genome assembly from NCBI, the Hymenoptera Genome Database (Munoz-Torres *et al.* 2011), or directly from authors (Johnson *et al.* 2013). To extract the UCE loci, we aligned our UCE bait sequences to each genome using a wrapper (phyluce_probe_run_multiple_lastzs_sqlite) around LASTZ with a less conservative setting of 80% for the 'coverage' and 'identity' parameters. We then used another script (phyluce_probe_slice_sequence_from_genomes) to slice out sequence corresponding to each UCE locus identified, along with 700 bp DNA flanking either side of the buffered UCE region (160 bp). After slicing, we used a script (phyluce_assembly_match_contigs_to_baits) to determine the detection or non-detection of UCE loci across lineages and to combine the results with those from sequenced taxa.

UCE DATA PROCESSING AND MATRIX GENERATION

The sequencing facility demultiplexed and converted raw data from BCL to FASTQ format. Using the FASTQ files, we cleaned and trimmed raw reads using ILLUMIPROCESSOR (Faircloth 2013), which is a wrapper around TRIMMOMATIC (Lohse *et al.* 2012; Del Fabbro *et al.* 2013), and we assembled reads *de novo* using a wrapper (phyluce_assembly_assemblo_trinity) around TRINITY v2013-02-25 (Grabherr *et al.* 2011). After assembly, we used a program (phyluce_assembly_match_contigs_to_baits) to identify individual UCE loci from the bulk of assembled contigs and to remove paralogs. We ran the script with default parameters and compared results using the following bait files as input (i) the principal hym-v2 bait file, containing bait sequences from all exemplar genomes; (ii) the ant-specific hym-v2 bait file, containing bait sequences from the ant genomes only and (iii) the hym-v1 bait file. The principal hym-v2 bait file produced the highest

Table 1. *In silico* capture results for UCEs and exons comparing the use of different bait set files. The files were used as input in the PHYLUCE v1.6 program *phyluce_assembly_match_contigs_to_probes*. For UCE capture we compared the hym-v1 bait file, the principal hym-v2 bait file and the ant-specific hym-v2 bait file. For exon capture we compared use of the synthesized bait sequences grouped by exon, the same grouped by gene, a set of complete exon sequences for 40 ant taxa, and the same but with only one exon per gene included ('gene2' in table). Note that we performed all *in vitro* enrichments using the ant-specific hym-v2 bait set, which includes 452 baits targeting 12 legacy phylogenetic markers

	All taxa			Ants only			Outgroups		
Bait file	Mean	Range	95 CI ±	Mean	Range	95 CI ±	Mean	Range	95 CI ±
uce-hym-v1	874	738–1002	10·6	879	751–1002	9·8	772	738–812	30·5
uce-hym-v2	2214	1445–2365	38·3	2249	1933–2365	18·6	1514	1445–1661	97·3
uce-hym-v2-ants	2205	1372–2357	39·5	2241	1931–2357	18·6	1476	1372–1623	103·4
exon-4t-exon	7·2	5–12	0·3	7·2	5–12	0·3	7·3	5–10	2·0
exon-4t-gene	8·8	6–11	0·3	8·8	7–11	0·3	8·3	6–10	1·7
exon-40t-gene	9·7	5–12	0·3	9·8	7–12	0·3	7·3	5–9	1·7
exon-40t-gene2	10·1	5–12	0·3	10·2	7–12	0·2	7·0	5–9	1·6

Table 2. Selected statistics for various alignment sets analysed in this study. The 'Matrix name' provides information on the filtering level for taxon occupancy ('F'). The 'Ants101T-F90-1238' alignment set was filtered to remove uninformative loci, loci with high rates of evolution, and loci with high levels of GC variance among taxa. The 'Ants101T-F90-1263' alignment set was filtered to remove loci exhibiting significant base composition heterogeneity among taxa

Matrix name	Loci	Length (bp)	$\bar{x}$ locus length (bp)	Locus range (bp)	Locus 95 CI (bp)	Inform sites (bp)	% missing data
Ants101T-F25	2523	1 508 121	597·8	180–1671	7·7	883 215	22·5
Ants101T-F50	2462	1 470 193	597·2	180–1671	7·8	866 954	21·3
Ants101T-F75	2304	1 381 459	599·6	180–1671	8·0	820 220	19·7
Ants101T-F90	1856	1 109 810	598·0	201–1671	8·8	667 208	17·5
Ants101T-F95	1122	665 218	592·9	201–1279	11·1	399 701	15·5
Ants101T-F98	435	261 335	600·8	201–1279	18·3	154 892	13·5
Ants101T-F100	19	10 947	576·2	340–1100	98·5	6429	11·9
Ants101T-F90-1238	1238	767 292	619·78	213–1671	10·4	452 953	16·4
Ants101T-F90-1263	1263	673 377	533·16	201–1671	8·6	408 893	17·1

per/sample capture results, so we used the set of UCE contigs identified with this file for subsequent processing steps (Table 1).

After UCE locus identification, we used two scripts (phyluce_assembly_get_match_counts and phyluce_assembly_get_fastas_from_match_counts) to combine the UCE contigs from the sequenced taxa with UCE contigs from 17 genome-enabled taxa (ants plus a few closely related outgroups) into a single, monolithic FASTA file. We aligned all loci individually using a wrapper (phyluce_align_seqcap_align) around MAFFT v7.130b (Katoh *et al.* 2002), and we trimmed alignments using a wrapper (phyluce_align_get_gblocks_trimmed_alignments_from_untrimmed) around GBLOCKS v0.91b (Castresana 2000; Talavera & Castresana 2007), which we ran with reduced stringency settings of 0·5, 0·5, 12 and 7 for the b1–4 settings, respectively. Finally, we used a script (phyluce_align_get_only_loci_with_min_taxa) to remove loci that had data for fewer than 90% of taxa (=91/101 taxa), and we used another script (phyluce_align_format_nexus_files_for_raxml) to generate a concatenated matrix from the resulting alignment set. We refer to the 90% filtered matrix and alignment set as 'Ants101T-F90'. Although we used the 90% filtered matrix for most analyses below, we also evaluated the number of loci present in matrices filtered to have 25, 50, 75, 95, 98 and 100% taxon occupancy (Table 2).

ANT PHYLOGENOMICS USING UCE LOCI

We conducted multiple phylogenomic analyses on the complete dataset to explore the ability of the new UCE data to resolve relationships among ant subfamilies and genera. We separate our analyses into three categories: (i) 'partitioning analyses', (ii) 'bias filtering' analyses and (iii) 'species tree' analyses. For all concatenated analyses, we used the maximum likelihood program RAxML v8 (Stamatakis 2014) and we performed a best tree plus rapid bootstrap search ('-f a' option) using GTR+Γ as the model of sequence evolution and 100 bootstrap replicates.

For the partitioning analyses, we analysed the complete Ants101T-F90 matrix three ways: (i) unpartitioned, (ii) partitioned by locus and (iii) partitioned using the hcluster algorithm in PartitionFinder v1.1.1 (Lanfear *et al.* 2012) (data pre-partitioned by locus). For the bias filtering analyses, we attempted to reduce possible negative effects caused by base composition heterogeneity, saturation and/or low information content. We performed three different concatenated analyses: (i) we converted the matrix to RY-coding; (ii) we removed loci that produced gene-trees with low mean bootstrap support (lowest 10%), high rates of evolution (highest 10%) and high levels of GC variance among taxa (highest 10%) and (iii) we removed all loci deviating significantly from base-composition homogeneity, as calculated with the program BaCoCa v1.1 (Kück & Struck 2014). We call the concatenated matrices for analyses 2 and 3 'Ants101T-F90-1238' and 'Ants101T-F90-1263', respectively, with the last number indicating the number of retained loci. For species tree estimation, we generated gene trees for all loci using RAxML, and then we input only the 500 'best' loci (best = highest mean gene-tree bootstrap scores) into the program ASTRAL v4.10.8 (Mirarab *et al.* 2014; Mirarab & Warnow 2015). We used only the 500 best loci to reduce bias from loci with low information content (see Meiklejohn *et al.* 2016). We calculated species tree support by performing 100 multi-locus bootstrap replicates (Seo 2008). For additional details on phylogenomic analyses see Appendix S2.

EXON EXTRACTION & MATRIX GENERATION

The synthesis of the ant-specific hym-v2 bait set included baits targeting 16 exons from 12 commonly sequenced nuclear genes. To extract these exons from the 84 enriched taxa and to combine the data with publicly available sequences, we performed the following steps: First, we created four different 'bait' files for extracting the loci from the complete pool of contigs. These bait files were: (i) all of the 452 synthesized bait sequences, grouped by exon; (ii) the same as (i), but grouped by gene; (iii) complete exon sequences for 40 different ant taxa and all targeted exons (P.S. Ward, unpublished data), grouped by gene; and (iv) the same as (iii), but with only one exon included per gene (*ArgK*, *CAD*, *LwRh* have multiple exons). Next, we tested the ability of the different bait files to extract exon sequences from the pool of contigs for each sample using a script (phyluce_assembly_match_contigs_to_baits). We ran the script using 80% for the 'min-coverage' and 'min-identity' parameters, and we created a duplicates file using the 'keep-duplicates' command. This duplicates file records which loci PHYLUCE removes due to possible paralogy. After comparing results for each of the bait files, we found that bait file (iv) performed the best across all taxa and across ants (Table 1). Thus, we used this set of identified contigs for subsequent steps. After extracting the exons, we used two scripts (phyluce_assembly_get_match_counts and phyluce_assembly_get_fastas_from_match_counts) to generate a monolithic FASTA file containing all taxa and exons.

For two loci (*EF1αF1* and *EF1αF2*) the above approach performed poorly because the copy variants were removed as paralogs. To add these loci to the dataset, we used the duplicates file generated by PHYLUCE to identify and manually extract the exons from the contigs file of each sample. We verified the identity of each *EF1α* copy using BLAST, and appended the sequences to the master FASTA file containing all data.

In addition to the exons, we used PHYLUCE to extract sequence data for the ribosomal RNA genes *18S* and *28S* from off-target reads. Specifically, we downloaded sequences of *18S* and *28S* from GenBank (*M. tahoensis*; AY703495.1 and AY703562.1), and we used a script (phyluce_assembly_match_contigs_to_barcodes) to slice out matching sequence from the pool of contigs for each enriched sample. We then manually checked the extracted sequences and appended each to the monolithic FASTA file. Although the bait set we synthesized does not specifically target these loci, they are present in multiple copies in the genome and are thus more likely to be present in off-target sequences than other less abundant genes.

After extracting the exon and rDNA loci from enriched samples, we used a script (phyluce_align_seqcap_align) to separate, align (with MAFFT) and trim sequences from individual loci using default settings. We used MESQUITE v3.0.3 (Maddison & Maddison 2016) to inspect each resulting alignment, and we removed introns and excessive flanking DNA. Next, we combined the exon and rDNA loci that we extracted from enriched samples with sequence data from published ant datasets. These included studies on all ants (Brady *et al.* 2006; Moreau & Bell 2013), and more focused studies on the ant subfamilies Dolichoderinae (Ward *et al.* 2010), Dorylinae (Brady *et al.* 2014), Myrmicinae (Ward *et al.* 2015), Formicinae (Blaimer *et al.* 2015), Ponerinae (Schmidt 2013), Amblyoponinae (Ward & Fisher 2016) and Pseudomyrmecinae (Chomicki, Ward & Renner 2015). For details on how the datasets were combined for analysis see Appendix S2.

PHYLOGENETIC ANALYSIS OF EXON DATA

To analyse the combined data matrix containing exon and rDNA loci from enriched and published samples, we pre-partitioned the matrix by gene and codon position and used PartitionFinder v1.1.1 to select the best partitioning scheme. We analysed the matrix in RAxML using a partitioned best tree plus rapid bootstrap search (GTR+Γ, 100 bootstrap replicates). To improve backbone support, we performed two additional analyses. First, we conducted a constraint analysis in RAxML ('-f a' and '-r' options), in which we used one of the UCE topologies (result from hcluster-partitioned analysis with genome-enabled taxa pruned) as a fixed backbone. For this search we used GTR+Γ as the model of sequence evolution and performed 100 rapid bootstrap replicates. In the second analysis, we created a new matrix in which we concatenated sequence data from the enriched exon and rDNA sequences with data from the 100 'best' performing UCE loci (best = highest mean gene-tree bootstrap scores). We analysed the resulting unpartitioned matrix in RAxML using GTR + CAT (selected to decrease computation time) as the model of sequence evolution and 100 bootstrap replicates. For all resulting trees, we modified the names of taxa to reflect the current state of ant nomenclature.

Results

BAIT DESIGN

We simulated an average of 2·64 M (95 CI ± 0·2 M) read pairs from exemplar hymenopteran lineages. On average, 3·0% (95 CI ± 0·85%) of these reads mapped to the *A. rosae* base genome (Table S5). After merging intervals, filtering short and masked loci, and determining presence/absence of loci across genomes, we identified 3010 conserved regions shared among *A. rosae* and all six exemplar lineages. We designed 5874 temporary baits targeting 2969 of these conserved regions. We aligned this temporary bait set to an average of 2197 non-duplicate conserved regions (95 CI ± 79·7%) across each exemplar genome, and after determining which loci we detected in the exemplar taxa, we output 2161 conserved loci identified among *A. rosae* and at least five of the exemplar genomes. We designed 27 495 baits targeting 2161 loci from the total of seven hymenopteran genomes. After filtering for duplicate and proximate loci and merging the older (hym-v1) and newer bait sets together, the final hymenopteran bait set contained 31 829 baits targeting 2590 conserved loci (948 from the hym-v1 design; 1642 from hym-v2 design). The ant-specific bait subset contained 9446 baits targeting 2524 conserved loci (893 from hym-v1 design; 1631 from hym-v2 design). In addition, we added 452 baits targeting 16 commonly sequenced exons to the ant-specific bait set, resulting in a final ant-specific bait set containing 9898 baits.

UCE ENRICHMENT RESULTS AND MATRIX STATISTICS

Across the 84 enriched samples, including four non-formicid outgroups, we captured an average of 2214 loci per taxon using the principal hym-v2 bait set file (Tables 1 and S6). Among ants, the average was slightly higher at 2249 loci and ranged from 1933 to 2356 loci (95 CI ± 170 loci). There was no obvious pattern of biased capture across ants. As expected, the number of captured loci was lower in the non-formicid outgroup taxa, which had a mean capture of 1514 loci (range 1445–1661 loci; 95 CI ± 198 loci). For all taxa, the average length of UCE contigs was 922 bp (Table S6; range 369–1322, 95 CI ± 330 bp) and the average coverage per contig was 39*x* (range 11–77*x*; 95 CI ± 27*x*). We also examined the number of loci captured using two alternative bait set files: (i) ant-specific hym-v2 baits (as opposed to the principal hym-v2 baits, as above), and (ii) the hym-v1 bait set (Faircloth *et al.* 2015). Using the ant-specific bait set file produced results similar to, but slightly lower than, using all baits (Tables 1 and S6; 2205 loci vs. 2214), with the mean decrease in the number of captured loci per taxon being most pronounced in outgroups (38 less in outgroups vs. 7 less in ants). Using the original bait file from hym-v1, we recovered an average of 874 loci across all taxa (range 738 to 1002 loci), far lower than the new v2 bait set.

Across the 38 genome-enabled taxa, we successfully extracted a mean of 2326 equivalent UCE loci (Table S1; range 1406–2326; 95 CI ± 95·2). For ant genomes only, the number of extracted loci was higher, with 2471 equivalent UCE loci (range 2414–2508; 95 CI ± 16·6). Although the number of extracted loci was significantly lower than the average in several hymenopteran genomes, there was no obvious phylogenetic bias in the results.

After extracting UCE contigs from the set of 84 enriched taxa and 17 genome-enabled taxa, we used the set of loci that we obtained with the principal Hymenoptera bait set file for all further analyses. We selected this set of loci because it included the highest number of loci, as compared to other capture sets. We aligned and trimmed the loci and then filtered them for varying levels of taxon occupancy (25, 50, 75, 90, 95, 98, 100%). A complete set of statistics for each occupancy level is given in Table 2. The primary alignment set used for phylogenomic analysis (Ants101T-F90) was filtered to have 90% complete taxon occupancy per locus, resulting in a concatenated matrix that included 1856 loci and 1 109 810 bp of aligned sequence data.

ANT PHYLOGENOMICS USING UCE LOCI

Considering major ant lineages, our results (Figs 1 and S4–S10) mostly confirm relationships found in previous studies based on legacy markers (Brady *et al.* 2006; Moreau *et al.* 2006; Moreau & Bell 2013), suggesting that our UCE data are informative and reliable. Compared to previous studies, the most notable results include recovery of Leptanillinae as the sister group to all other ants (100% support in all analyses; Martialinae not included); Myrmicinae as the sister group to Ectatomminae + Heteroponerinae (100% support in concatenated analyses; 67% support in ASTRAL analysis); and complete resolution among the six myrmicine tribes, with Myrmicini + Pogonomyrmecini sister to [Stenammini + [Crematogastrini + [Solenopsidini + Attini]]]. Despite our use of genome-scale data, we could not reliably resolve relationships among poneroid subfamilies, except for Apomyrminae + Amblyoponinae. For a more detailed description of the phylogenetic results see Appendix S2.

EXTRACTION AND ANALYSIS OF EXON DATA

We successfully enriched, sequenced and extracted a set of 12 commonly sequenced genes (16 exons total) from all lineages. Using the custom 'bait' file targeting only one exon per gene, we recovered a mean of 10·1 genes from each enriched sample (Table 1), with the mean slightly lower in non-ant taxa compared to ants only (7·0 vs. 10·2 genes). After manually adding genes that had been removed by PHYLUCE, the mean number of recovered loci increased to 11·7 genes per taxon across all samples (8·3 genes for non-ants, 11·9 for ants). At the level of individual exons, we successfully recovered a mean of 15·5 exons (out of 16 total) per taxon (15·8 exons for ants, 10·0 for non-ants). For the ribosomal genes, which we skimmed from off-target reads, we successfully extracted each gene from 83 of 84 taxa. After combining all extracted data, the final dataset for enriched taxa included complete or partial sequence data for an average of 17·5 out of 18 loci (16 exons plus the two ribosomal genes) per taxon across all samples (17·7 for ants, 12 for non-ants).

Following extraction, we combined the exon + rDNA data with similar data from nine published studies on ants. After removing duplicate taxa and aligning and trimming the data, the combined data matrix included 1060 terminals, 15 nuclear genes (all loci mentioned above plus *Enolase*) and 11 406 bp of aligned sequence data, including 4678 informative sites. Phylogenetic results revealed that all enriched taxa were placed in sensible positions within the ant phylogeny (Figs S1–S3). For 34 of the 84 enriched samples, the combined dataset included sequence data from two conspecific samples, one from enriched data and one from published Sanger data, and in all these cases, the enriched sample was correctly placed as closely related to its conspecific duplicate. These results confirm that the extracted exon data can be reliably recovered along with the enriched UCE loci. Although we do not discuss the broader phylogenetic results, a few novel and/or interesting findings include paraphyly of *Proceratium*, *Prolasius* and *Heteroponera* and monophyly of *Tetraponera* and New World army ants.

Discussion

Sequence capture of conserved genomic loci (UCEs) has emerged as a revolutionary tool for efficiently investigating the evolutionary history of organisms (Faircloth *et al.* 2012, 2015; Blaimer *et al.* 2015; Branstetter *et al.* 2016b; Glenn & Faircloth 2016). Here, we designed and tested an enhanced UCE bait set for performing sequence capture in Hymenoptera, with a particular focus on ants.

We demonstrated that the new bait design approach improves upon the original design of Faircloth *et al.* (2015) in several important ways. First, by changing the workflow by

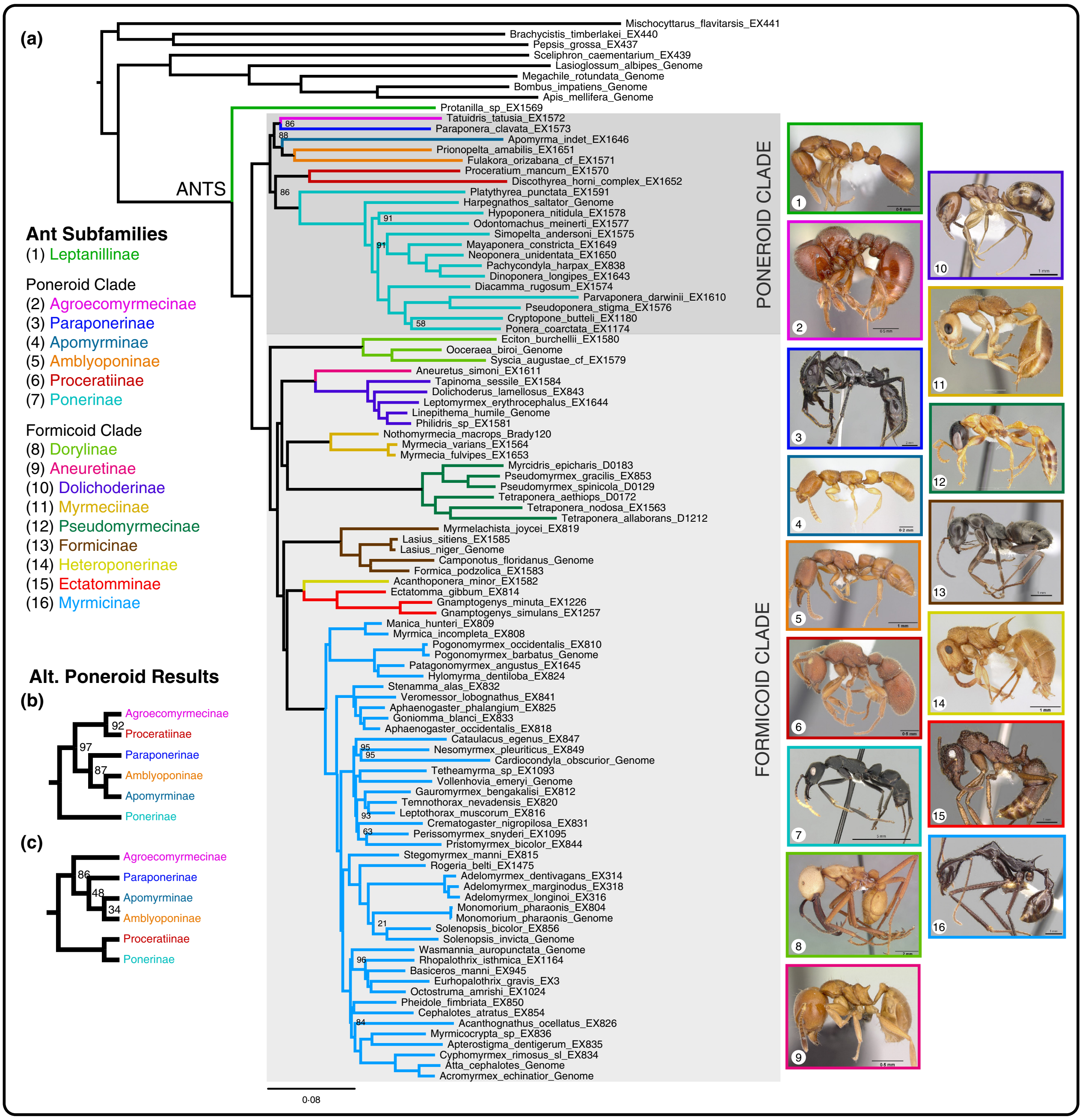
(a)
Mischocyttarus_flavitarsis_EX441
Brachycistis_timberlakei_EX440
Pepsis_grossa_EX437
Sceliphron_caementarium_EX439
Lasioglossum_albipes_Genome
Megachile_rotundata_Genome
Bombus_impatiens_Genome
Apis_mellifera_Genome
ANTS
Protanilla_sp_EX1569
Tatuidris_tatusia_EX1572
Paraponera_clavata_EX1573
Apomyrma_indet_EX1646
Prionopelta_amabilis_EX1651
Fulakora_orizabana_cf_EX1571
Proceratium_mancum_EX1570
Discothyrea_horni_complex_EX1652
Platythyrea_punctata_EX1591
Harpegnathos_saltator_Genome
Hypoponera_nitidula_EX1578
Odontomachus_meinerti_EX1577
Simopelta_andersoni_EX1575
Mayaponera_constricta_EX1649
Neoponera_unidentata_EX1650
Pachycondyla_harpax_EX838
Dinoponera_longipes_EX1643
Diacamma_rugosum_EX1574
Parvaponera_darwinii_EX1610
Pseudoponera_stigma_EX1576
Cryptopone_butteli_EX1180
Ponera_coarctata_EX1174
86
88
86
91
91
58
PONEROID CLADE
Eciton_burchellii_EX1580
Ooceraea_biroi_Genome
Syscia_augustae_cf_EX1579
Aneuretus_simoni_EX1611
Tapinoma_sessile_EX1584
Dolichoderus_lamellosus_EX843
Leptomyrmex_erythrocephalus_EX1644
Linepithema_humile_Genome
Philidris_sp_EX1581
Nothomyrmecia_macrops_Brady120
Myrmecia_varians_EX1564
Myrmecia_fulvipes_EX1653
Myrcidris_epicharis_D0183
Pseudomyrmex_gracilis_EX853
Pseudomyrmex_spinicola_D0129
Tetraponera_aethiops_D0172
Tetraponera_nodosa_EX1563
Tetraponera_allaborans_D1212
Myrmelachista_joycei_EX819
Lasius_sitiens_EX1585
Lasius_niger_Genome
Camponotus_floridanus_Genome
Formica_podzolica_EX1583
Acanthoponera_minor_EX1582
Ectatomma_gibbum_EX814
Gnamptogenys_minuta_EX1226
Gnamptogenys_simulans_EX1257
Manica_hunteri_EX809
Myrmica_incompleta_EX808
Pogonomyrmex_occidentalis_EX810
Pogonomyrmex_barbatus_Genome
Patagonomyrmex_angustus_EX1645
Hylomyrma_dentiloba_EX824
Stenamma_alas_EX832
Veromessor_lobognathus_EX841
Aphaenogaster_phalangium_EX825
Goniomma_blanci_EX833
Aphaenogaster_occidentalis_EX818
Cataulacus_egenus_EX847
Nesomyrmex_pleuriticus_EX849
Cardiocondyla_obscurior_Genome
Tetheamyrma_sp_EX1093
Vollenhovia_emeryi_Genome
Gauromyrmex_bengakalisi_EX812
Temnothorax_nevadensis_EX820
Leptothorax_muscorum_EX816
Crematogaster_nigropilosa_EX831
Perissomyrmex_snyderi_EX1095
Pristomyrmex_bicolor_EX844
Stegomyrmex_manni_EX815
Rogeria_belti_EX1475
Adelomyrmex_dentivagans_EX314
Adelomyrmex_marginodus_EX318
Adelomyrmex_longinoi_EX316
Monomorium_pharaonis_EX804
Monomorium_pharaonis_Genome
Solenopsis_bicolor_EX856
Solenopsis_invicta_Genome
Wasmannia_auropunctata_Genome
Rhopalothrix_isthmica_EX1164
Basiceros_manni_EX945
Eurhopalothrix_gravis_EX3
Octostruma_amrishi_EX1024
Pheidole_fimbriata_EX850
Cephalotes_atratus_EX854
Acanthognathus_ocellatus_EX826
Myrmicocrypta_sp_EX836
Apterostigma_dentigerum_EX835
Cyphomyrmex_rimosus_sl_EX834
Atta_cephalotes_Genome
Acromyrmex_echinatior_Genome
95
95
93
63
21
96
84
0·08
FORMICOID CLADE
Ant Subfamilies
(1) Leptanillinae
Poneroid Clade
(2) Agroecomyrmecinae
(3) Paraponerinae
(4) Apomyrminae
(5) Amblyoponinae
(6) Proceratiinae
(7) Ponerinae
Formicoid Clade
(8) Dorylinae
(9) Aneuretinae
(10) Dolichoderinae
(11) Myrmeciinae
(12) Pseudomyrmecinae
(13) Formicinae
(14) Heteroponerinae
(15) Ectatomminae
(16) Myrmicinae
Alt. Poneroid Results
(b)
Agroecomyrmecinae
92
Proceratiinae
97
Paraponerinae
87
Amblyoponinae
Apomyrminae
Ponerinae
(c)
Agroecomyrmecinae
86
Paraponerinae
48
Apomyrminae
34
Amblyoponinae
Proceratiinae
Ponerinae
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Fig. 1. Phylogeny of ant subfamilies, derived from genome-scale data. (a) ML tree from hcluster-partitioned analysis of 1856 UCE loci. Trees (b–c) pruned to show alternative relationships for poneroid subfamilies only. (b) ML tree from RY-coded analysis of all loci. (c) ASTRAL species tree generated from the 500 loci with the highest mean gene-tree bootstrap scores. All nodes with 100% support unless indicated. Branches colour coded by ant subfamily. See also Figs S4–S10. Ant images from www.antweb.org.

which we identified conserved loci, we increased the number of targeted loci by 1080. Second, by including a greater number of genomes in locus identification, bait design and bait synthesis, we increased the capture success of UCE loci across ants (mean 2249 loci from enriched samples) and Hymenoptera (mean 2326 loci from 38 genome-enabled taxa). In both cases, there was no obvious pattern of biased capture performance across taxa. Because of the high average capture rate, we were able to generate a phylogenomic matrix for ants that included more loci and fewer missing data (1856 loci and 17·5% missing data in the Ants101T-F90 matrix) than previously possible using the hym-v1 bait set. Third, by targeting loci from the original bait set and legacy markers, data generated with the new bait set are combinable with published datasets. The power of this feature was exemplified by our ability to combine exon data extracted from enriched taxa with data from multiple studies. In so doing, we created the most inclusive ant phylogeny to date. It is not yet known whether data from our bait set will be combinable with data from other approaches (e.g. transcriptomes), but there is likely to be some overlap because approximately 61% of the UCE loci we captured include some coding sequence (M.G. Branstetter, unpublished data).

Finally, our bait set is customizable. By designing the principal hymenopteran bait set from seven different genomes spread across Hymenoptera, researchers can use the principal bait set in its entirety to work across all Hymenoptera or they can subset the principal bait set to focus on a particular group within Hymenoptera. In our case, we synthesized an ant-specific bait set that included baits from the phylogenetically distant ants *H. saltator* and *A. cephalotes*, while excluding baits that were unnecessary to answer our research questions. As new hymenopteran genomes become available, it will also be possible to design new customized baits by aligning our baits to these genomes. The tools for performing such customization are available as part of the updated UCE design workflow (Faircloth in press; tutorial available at http://phyluce.readthedocs.io/en/latest/index.html).

It remains unclear how many loci are necessary to resolve the majority of phylogenetic relationships, but we believe that having a greater number of loci to work with is beneficial. A major component of modern phylogenomic analysis is the ability to remove, or 'filter', loci that appear to have undesirable qualities, such as missing data, base composition bias, or saturation (Borowiec *et al.* 2015; Fernández, Edgecombe & Giribet 2016; Meiklejohn *et al.* 2016). Consequently, having more loci gives researchers greater flexibility to remove potentially problematic loci, resulting in more robust analyses and results. In this study, we filtered loci for missing data, low information content and base composition heterogeneity, and we were still able to include over 1000 loci in most analyses. Also, while some loci might be problematic in all datasets, it is likely that the set of loci that need to be filtered will vary depending on taxonomic depth and focal lineage. Thus, having a broader range of loci to choose from provides greater phylogenetic power.

Our phylogenomic results indicate that the data generated by the new bait set are phylogenetically informative within ants and close relatives and have the potential to resolve previously intractable problems. Of greatest significance, we found Myrmicinae to be sister to the ectaheteromorph clade, and we recovered unequivocal support for novel relationships among myrmicine tribes. Relationships among poneroid subfamilies were not well resolved, however, even with genome-scale data. Improving results in this part of the ant tree will likely require improved taxon sampling to break up long branches.

Authors' contributions

All authors conceived the ideas and designed methodology; M.G.B. and B.C.F. collected the data; M.G.B. and B.C.F. analysed the data; M.G.B. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgements

We thank Mac Pierce, Theo Sumnicht and Eric Youngerman for laboratory assistance and the Brazelton Lab (University of Utah) for use of their Linux server. We thank Stephen Richards, Masatsugu Hatakeyama, Evgenii Konorov, Sasha Mikheyev, Catherine Linnen, Oliver Niehuis and Scott Geib for allowing us to use unpublished genome data. Brian Dalley and Brett Milash from the University of Utah Genomics and Bioinformatics Core provided assistance with sequencing and bioinformatics. Phylogenetic analyses were performed using either the CIPRES Science Gateway (Miller, Pfeiffer & Schwartz 2010) or the Smithsonian Institution's high-performance computer cluster (Hydra). We thank Rosa Fernándes, Samantha Price, and one anonymous reviewer for helpful comments on the manuscript. This study was supported by National Science Foundation grants DEB-1354739 & DEB-1354996 (Project ADMAC: Ant Diversity of the MesoAmerican Corridor).

Data accessibility

All bait set files used in this study are available from Dryad (Branstetter *et al.* 2017; https://doi.org/10.5061/dryad.89n87), and from FigShare (https://doi.org/10.6084/m9.figshare.4630375.v1), where we will maintain updated/improved versions. The PHYLUCE software package is available from GitHub (https://github.com/faircloth-lab/phyluce). Raw sequence reads and UCE contig assemblies are available from the NCBI Sequence Read Archive and GenBank respectively (NCBI BioProject PRJNA360290). Additional data, including alignments, trees, UCE contigs and tables are also available on Dryad (https://doi.org/10.5061/dryad.89n87).

References

Blaimer, B.B., Brady, S.G., Schultz, T.R., Lloyd, M.W., Fisher, B.L. & Ward, P.S. (2015) Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evolutionary Biology*, **15**, 271.

Blaimer, B.B., LaPolla, J.S., Branstetter, M.G., Lloyd, M.W. & Brady, S.G. (2016a) Phylogenomics, biogeography and diversification of obligate mealy-bug-tending ants in the genus *Acropyga*. *Molecular Phylogenetics & Evolution*, **102**, 20–29.

Blaimer, B.B., Lloyd, M.W., Guillory, W.X. & Brady, S.G. (2016b) Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS ONE*, **11**, e0161531.

Borowiec, M.L., Lee, E.K., Chiu, J.C. & Plachetzki, D.C. (2015) Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics*, **16**, 987.

Brady, S.G., Schultz, T.R., Fisher, B.L. & Ward, P.S. (2006) Evaluating alternative hypotheses for the early evolution and diversification of ants. *Proceedings of the National Academy of Sciences USA*, **103**, 18172–18177.

Brady, S.G., Fisher, B.L., Schultz, T.R. & Ward, P.S. (2014) The rise of army ants and their relatives: diversification of specialized predatory doryline ants. *BMC Evolutionary Biology*, **14**, 93.

Branstetter, M.G., Longino, J.T., Reyes-López, J., Schultz, T.R. & Brady, S.G. (2016a) Into the tropics: phylogenomics and evolutionary dynamics of a contrarian clade of ants. *bioRxiv*, 1–52. doi: 10.1101/039966.

Branstetter, M.G., Danforth, B.N., Pitts, J.P., Faircloth, B.C., Ward, P.S., Buffington, M.L., Gates, M.G., Kula, R.R. & Brady, S.G. (2016b) Phylogenomic analysis of ants, bees, and stinging wasps: improved taxon sampling enhances understanding of hymenopteran evolution. *bioRxiv*, 1–40. doi: 10.1101/068957.

Branstetter, M.G., Longino, J.T., Ward, P.S. & Faircloth, B.C. (2017) Data from: Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Dryad Digital Repository*. Available at: http://dx.doi.org/10.5061/dryad.89n87.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Chomicki, G., Ward, P.S. & Renner, S.S. (2015) Macroevolutionary assembly of ant/plant symbioses: *Pseudomyrmex* ants and their ant-housing plants in the Neotropics. *Proceedings of the Royal Society B*, **282**, 20152200.

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K. & Glenn, T.C. (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, **8**, 783–786.

Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F.M. (2013) An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS ONE*, **8**, e85024.

Faircloth, B.C. (2013) Illumiprocessor: A trimmomatic wrapper for parallel adapter and quality trimming. Available at: https://doi.org/10.6079/J9ILL.

Faircloth, B.C. (in press). Identifying conserved genomic elements and designing universal probe sets to enrich them. *Methods in Ecology and Evolution*. doi: 10.1111/2041-210X.12754.

Faircloth, B.C. (2016) PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, **32**, 786–788.

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T. & Glenn, T.C. (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Faircloth, B.C., Sorenson, L., Santini, F. & Alfaro, M.E. (2013) A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *PLoS ONE*, **8**, e65923.

Faircloth, B.C., Branstetter, M.G., White, N.D. & Brady, S.G. (2015) Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular Ecology Resources*, **15**, 489–501.

Fernández, R., Edgecombe, G.D. & Giribet, G. (2016) Exploring phylogenetic relationships within Myriapoda and the effects of matrix composition and occupancy on phylogenomic reconstruction. *Systematic Biology*, **65**, 871–889.

Glenn, T.C. & Faircloth, B.C. (2016) Capturing Darwin's dream. *Molecular Ecology Resources*, **16**, 1051–1058.

Grabherr, M.G., Haas, B.J., Yassour, M., *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Harris, R.S. (2007) *Improved pairwise alignment of genomic DNA*. Ph.D. thesis, The Pennsylvania State University, State College, PA, USA.

Harvey, M.G., Smith, B.T., Glenn, T.C., Faircloth, B.C. & Brumfield, R.T. (2016) Sequence capture versus restriction site associated DNA sequencing for phylogeography. *Systematic Biology*, **65**, 910–924.

Huang, W., Li, L., Myers, J.R. & Marth, G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Johnson, B.R., Borowiec, M.L., Chiu, J.C., Lee, E.K., Atallah, J. & Ward, P.S. (2013) Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. *Current Biology*, **23**, 1–5.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Kück, P. & Struck, T.H. (2014) BaCoCa – A heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Molecular Phylogenetics and Evolution*, **70**, 94–98.

Lanfear, R., Calcott, B., Ho, S.Y.W. & Guindon, S. (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, **29**, 1695–1701.

Lemmon, E.M. & Lemmon, A.R. (2013) High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 19.1–19.23.

Li, H., Handsaker, B., Wysoker, A., *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lohse, M., Bolger, A.M., Nagel, A., Fernie, A.R., Lunn, J.E., Stitt, M. & Usadel, B. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, W622–W627.

Lunter, G. & Goodson, M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Maddison, W.P. & Maddison, D.R. (2016) Mesquite: a modular system for evolutionary analysis. Version 3.03. Available at: http://mesquiteproject.org (accessed 6 March 2015).

Manthey, J.D., Campillo, L.C., Burns, K.J. & Moyle, R.G. (2016) Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: *Piranga*). *Systematic Biology*, **65**, 640–650.

McCormack, J.E., Tsai, W.L.E. & Faircloth, B.C. (2015) Sequence capture of ultraconserved elements from bird museum specimens. *Molecular Ecology Resources*, **16**, 1189–1203.

McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T. & Glenn, T.C. (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, **22**, 746–754.

McCormack, J.E., Harvey, M.G., Faircloth, B.C., Crawford, N.G., Glenn, T.C. & Brumfield, R.T. (2013a) A phylogeny of birds based on over 1,500 Loci collected by target enrichment and high-throughput sequencing. *PLoS ONE*, **8**, e54848.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C. & Brumfield, R.T. (2013b) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Kimball, R.T. & Braun, E.L. (2016) Analysis of a rapid evolutionary radiation using ultraconserved elements (UCEs): evidence for a bias in some multispecies coalescent methods. *Systematic Biology*, **65**, 612–627.

Miller, M.A., Pfeiffer, W. & Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 November, 2010, New Orleans, LA, USA, pp. 1–8.

Mirarab, S. & Warnow, T. (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatic*, **31**, i44–i52.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S. & Warnow, T. (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548.

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

Munoz-Torres, M.C., Reese, J.T., Childers, C.P., Bennett, A.K., Sundaram, J.P., Childs, K.L., Anzola, J.M., Milshina, N. & Elsik, C.G. (2011) Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Research*, **39**, D658–D662.

Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Schmidt, C. (2013) Molecular phylogenetics of ponerine ants (Hymenoptera: Formicidae: Ponerinae). *Zootaxa*, **3647**, 201–250.

Seo, T.K. (2008) Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, **25**, 960–971.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Starrett, J., Derkarabetian, S., Hedin, M., Bryson, R.W., McCormack, J.E. & Faircloth, B.C. (2016) High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. *Molecular Ecology Resources*, doi: 10.1111/1755-0998.12621. [Epub ahead of print].

Talavera, G. & Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564–577.

Ward, P.S. (2014) The phylogeny and evolution of ants. *Annual Review of Ecology, Evolution, and Systematics*, **45**, 23–43.

Ward, P.S. & Fisher, B.L. (2016) Tales of dracula ants: the evolutionary history of the ant subfamily Amblyoponinae (Hymenoptera: Formicidae). *Systematic Entomology*, **41**, 683–693.

Ward, P.S., Brady, S.G., Fisher, B.L. & Schultz, T.R. (2010) Phylogeny and biogeography of dolichoderine ants: effects of data partitioning and relict taxa on historical inference. *Systematic Biology*, **59**, 342–362.

Ward, P.S., Brady, S.G., Fisher, B.L. & Schultz, T.R. (2015) The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae). *Systematic Entomology*, **40**, 61–81.

Yeates, D.K., Meusemann, K., Trautwein, M., Wiegmann, B. & Zwick, A. (2016) Power, resolution and bias: recent advances in insect phylogeny driven by the genomic revolution. *Current Opinion in Insect Science*, **13**, 16–23.

Received 19 November 2016; accepted 14 January 2017
Handling Editor: Samantha Price

Supporting Information

Details of electronic Supporting Information are provided below.

Fig. S1. Comprehensive ant phylogeny (1060 terminals) inferred with RAxML (partitioned).

Fig. S2. Comprehensive ant phylogeny (1060 terminals) inferred with RAxML (constrained and partitioned).

Fig. S3. Comprehensive ant phylogeny (1060 terminals) inferred with RAxML.

Appendix S1. Supplemental tables.

Appendix S2. Supplemental methods, table captions, figures and references.