MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12328

Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera

BRANT C. FAIRCLOTH,*,† MICHAEL G. BRANSTETTER,‡ NOOR D. WHITE§,¶ and SEÁN G. BRADY‡
*Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA, †Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA, ‡Department of Entomology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA, §Department of Biology, University of Maryland, College Park, MD 20742, USA, ¶Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA

Abstract

Gaining a genomic perspective on phylogeny requires the collection of data from many putatively independent loci across the genome. Among insects, an increasingly common approach to collecting this class of data involves transcriptome sequencing, because few insects have high-quality genome sequences available; assembling new genomes remains a limiting factor; the transcribed portion of the genome is a reasonable, reduced subset of the genome to target; and the data collected from transcribed portions of the genome are similar in composition to the types of data with which biologists have traditionally worked (e.g. exons). However, molecular techniques requiring RNA as a template, including transcriptome sequencing, are limited to using very high-quality source materials, which are often unavailable from a large proportion of biologically important insect samples. Recent research suggests that DNA-based target enrichment of conserved genomic elements offers another path to collecting phylogenomic data across insect taxa, provided that conserved elements are present in and can be collected from insect genomes. Here, we identify a large set (n = 1510) of ultraconserved elements (UCEs) shared among the insect order Hymenoptera. We used *in silico* analyses to show that these loci accurately reconstruct relationships among genome-enabled hymenoptera, and we designed a set of RNA baits (n = 2749) for enriching these loci that researchers can use with DNA templates extracted from a variety of sources. We used our UCE bait set to enrich an average of 721 UCE loci from 30 hymenopteran taxa, and we used these UCE loci to reconstruct phylogenetic relationships spanning very old (≥220 Ma) to very young (≤1 Ma) divergences among hymenopteran lineages. In contrast to a recent study addressing hymenopteran phylogeny using transcriptome data, we found ants to be sister to all remaining aculeate lineages with complete support, although this result could be explained by factors such as taxon sampling. We discuss this approach and our results in the context of elucidating the evolutionary history of one of the most diverse and speciose animal orders.

Keywords: arthropods, baits, conserved sequence, Hymenoptera, probes, target enrichment, ultraconserved elements

Received 31 May 2014; revision received 31 August 2014; accepted 5 September 2014

Introduction

The insect order Hymenoptera (ants, bees, wasps and sawflies) is one of the most diverse animal orders (Grimaldi & Engel 2005), including at least 125 000 species (Gaston 1991; LaSalle & Gauld 1993; Gaston *et al.* 1996; Sharkey 2007) and comprising approximately 8% of all described animals (Davis *et al.* 2010). In addition to being hyperdiverse, hymenopterans are also incredibly abundant: ants alone occur in almost all terrestrial habitats and may constitute 15–20% of animal biomass in tropical rainforests. Other aculeate groups such as vespid wasps (hornets and yellow jackets) and stingless honey bees may account for an additional 15–20% (Fittkau & Klinge 1973). The ecological roles held by hymenopterans are diverse and include predator, prey, scavenger, parasite, ectosymbiont and mutualist. Species within the order also play critical roles in worldwide pollination of agricultural crops and natural vegetation (Roubik 1995; Kevan 1999; Michener 2007), tropical forest dynamics (Roubik 1989; Levey & Byrne

Correspondence: Brant C. Faircloth, Fax: 225-578-2597;
E-mail: brant@faircloth-lab.org

1993; Dalling *et al.* 1998), ecosystem services (Kremen *et al.* 2002; Del Toro *et al.* 2012) and biological control of pest insects (Quicke 1997). Outside of their biological importance, hymenopteran taxa are models for understanding the evolution and genetic basis of social behaviour (Bourke & Franks 1995; Smith *et al.* 2008; Bradley *et al.* 2009; Johnson & Linksaver 2010; Howard & Thorne 2010; Wang *et al.* 2013), mechanisms of sex determination (Hunt & Page 1994; Beye *et al.* 1994; Page *et al.* 2002), evolution of adaptive specialization (Mueller *et al.* 2005; Schultz & Brady 2008; Mueller & Rabeling 2008) and origins and maintenance of biodiversity (reviewed in LaSalle & Gauld 1993).

Given their ubiquity, diversity, biological significance and importance to ecological and agricultural systems, resolving evolutionary relationships among Hymenoptera is critical – from the deepest splits (*c.* 220–300 Ma) within the Hymenoptera tree (Grimaldi & Engel 2005; Ronquist *et al.* 2012) to moderately deep divergences (*c.* 120–60 Ma) comprising key events in the evolution of both the ecologically dominant ants [the 'Dynastic-Succession' hypothesis of (Wilson & Hölldobler 2005)] and pollinating bees (Danforth *et al.* 2013) to the very shallow divergences among lineages that may be undergoing ecological (Savolainen & Vepsäläinen 2003) or symbiont-driven speciation (Mehdiabadi *et al.* 2012). Prior molecular phylogenetic studies have made significant advances towards resolving the relationships between higher-level taxonomic groups (Sharkey 2007; Pilgrim *et al.* 2008; Heraty *et al.* 2011; Debevec *et al.* 2012; Klopfstein *et al.* 2013) and elucidating taxonomic relationships among species at shallower levels (reviewed in Moreau 2009; Danforth *et al.* 2013). However, these studies have been limited to analysing a relatively small number of nuclear or mitochondrial loci (e.g. Brady *et al.* 2006; Danforth *et al.* 2006; Sharanowski *et al.* 2010) that sample a small fraction of the genome.

Phylogenomic projects, such as the 1KITE initiative (http://www.1kite.org), seek to remedy this shortfall by identifying orthologous loci from widespread transcriptome sequencing. Although this approach has proven effective within Hymenoptera (Johnson *et al.* 2013), RNA-based techniques, on their own, limit the source materials useable for phylogenetic inference to fresh or properly preserved tissue (e.g. tissues stored in liquid nitrogen or RNAlater). This restriction leaves the majority of insect specimens unusable, especially those materials found in museum collections, posing a significant challenge for studies requiring rarely collected species. Thus, a significant challenge that remains for hymenopteran phylogenetics is to identify a large suite of universal markers that can be applied to samples stored with minimal preservation while maintaining the capability to elucidate relationships among lineages across a diversity of timescales.

Recent research among vertebrates has shown that target enrichment of highly conserved genomic sequences or 'ultraconserved elements' (UCEs; Faircloth *et al.* 2012) provides one mechanism for meeting this challenge. UCEs are an ideal marker for phylogenetic studies as a result of their ubiquity among taxonomic groups (Siepel *et al.* 2005), low paralogy (Derti *et al.* 2006) and low saturation (McCormack *et al.* 2012). While we still do not understand the evolutionary forces driving the conservation of UCEs (Harmston *et al.* 2013) or their biological function (Bejerano *et al.* 2004; Sandelin *et al.* 2004; Ahituv *et al.* 2007), target enrichment of UCE loci has been used to investigate several outstanding phylogenetic questions at 'deep' timescales across diverse groups of vertebrate taxa (Crawford *et al.* 2012; McCormack *et al.* 2013; Faircloth *et al.* 2013). The technique is also useful for understanding shallower, population-level events including recent divergences (Smith *et al.* 2014). When combined with massively parallel sequencing, the scalability of the UCE approach allows researchers to parallelize the collection of data from hundreds or thousands of orthologous loci across hundreds of taxa using stable DNA inputs in a single sequencing run; reduce the data analysis burden relative to what is required for the sequencing, assembly and alignment of multiple genomes; and conduct studies at a reasonable cost per individual.

Although enriching conserved loci resolves relationships among vertebrates, the utility of this approach among other animals is unknown. Here, we report the identification of a suite of *c.* 1500 UCE loci useful for inferring phylogenetic relationships across the entire Hymenoptera order. We used an *in silico* analysis to show that UCE loci recover the expected relationships among extant, genome-enabled, hymenopteran taxa with high support. We then synthesized a bait (i.e. probe) set for targeted enrichment of UCE loci, and we used the bait set to enrich an average of 721 loci among 30 sequence-tagged genomic libraries prepared from a diverse group of hymenopteran DNA sources, some of which were minimally preserved in ethanol for more than 12 years (Table S1, Supporting information). Using contigs assembled from massively parallel sequencing reads of these enriched libraries, we inferred the evolutionary relationships among hymenopteran taxa spanning very deep (≥220 Ma; estimated age of crown-group Hymenoptera; Grimaldi & Engel 2005; Ronquist *et al.* 2012) to very shallow (≤1 Ma; estimated age of included *Nasonia* species; Werren *et al.* 2010) divergences, and we discuss our findings relative to both phylogenomic and traditional efforts to resolve the hymenopteran phylogeny.

Materials and methods

Identification of UCEs

To identify a large set of UCE loci shared among Hymenoptera, we used LASTZ (Harris 2007) and programs from the UCE-PROBE-DESIGN package (UPDP) (https://github.com/faircloth-lab/uce-probe-design). We aligned repeat-masked (Smit *et al*. 1996-2010) genome assemblies of *Apis mellifera* (*apiMel4*; Honeybee Genome Sequencing Consortium 2006) and *Nasonia vitripennis* (*nasVit2*; Werren *et al.* 2010) using LASTZ. Following sequence alignment, we used rename_maf.py from UPDP to annotate the resulting multiple alignment format (MAF) lines with each taxon name. Following annotation, we used summary.py to search the resulting MAF file for aligned regions longer than 40 base pairs that were 100% conserved. We identified 2906 conserved regions meeting these criteria, and we filtered these regions for duplicate hits using an additional LASTZ alignment of conserved regions back to themselves (all-to-all) followed by removal of matches that were more than 80% identical over 50% of their length. After removing these duplicate-like regions, we output a file of 1555 nonduplicated UCE loci, and we checked for detection of these loci in two additional hymenopteran genome assemblies (*Atta cephalotes*, *Solenopsis invicta*; Table S2, Supporting information) by aligning the conserved regions to the assemblies using LASTZ, requiring 80% sequence identity over 80% of the nonduplicate UCE locus length. Approximately 1000–1300 of these UCE loci were conserved across the hymenopteran genome assemblies we checked, suggesting that the suite of nonduplicated, highly conserved loci we identified were also conserved in other hymenopteran lineages.

Based on this positive result, we sliced all of the nonduplicate UCE regions from the *nasVit2* genome sequence using match coordinates (as Browser Extensible Data or BED files) output by LASTZ, and we buffered shorter UCE regions to 180 bp by including an equal amount of 5′ and 3′ flanking sequence from the *nasVit2* genome assembly. This buffering process allowed us to tile 120 nucleotide enrichment baits across the desired target regions at 2X tiling density (i.e. baits overlap by 60 bp; Tewhey *et al.* 2009) using py_tiler.py from the UPDP. This program also removed any resulting baits containing ambiguous base calls, having a large proportion (>25%) of repetitive sequence or having a high GC content (>70%). We screened the resulting bait sequences against themselves to remove duplicate baits from the set that sometimes resulted from slicing longer, unique UCE loci into smaller, 120 nucleotide chunks. We refer to this final set of baits as the 'UCE bait set' below.

In silico test of UCEs

We performed an *in silico* test of the ability of the UCE baits and their target UCE loci to resolve the phylogeny of Hymenoptera by aligning the UCE bait set to 14 hymenopteran genome assemblies downloaded from NCBI (Table S2, Supporting information) using a parallel wrapper around LASTZ (run_multiple_lastzs_sqlite.py) from the PHYLUCE (https://github.com/faircloth-lab/phyluce) package. Although genome assemblies exist for additional hymenopteran taxa, we were not granted permission to include these data in our analyses. Following alignment, we sliced the UCE loci from each genome and retained ±1000 bp of flanking sequence to the 5′ and 3′ end of each UCE using slice_sequence_from_genomes2.py. This program makes a first pass at removing duplicate hits during the slicing process. After slicing, and to identify assembled contigs representing UCE loci from each species using the standard PHYLUCE pipeline, we aligned species-specific UCE slices to a FASTA file of all enrichment baits using match_contigs_to_loci.py from the PHYLUCE package. This program implements the matching process using LASTZ and ensures that matches are 80% identical over 80% of their length. This program also screens and removes apparent duplicate contigs or contigs that are hit by baits targeting more than one UCE locus. After screening and removing nontarget and duplicated or misassembled contigs, the program creates a relational database containing two tables – one that holds the status of each UCE locus in each taxon (detected/nondetected) and another that maps the contig names generated by the assembler to the names of the corresponding UCE locus across all taxa.

We created a file containing the names of 14 genome-enabled taxa (Table S2, Supporting information), and we input this list to an additional program (get_match_counts.py) that queries the relational database described above to generate a list of UCE loci shared among taxa. We input the list of loci generated by this program to another program (get_fastas_from_match_counts.py) to create a monolithic FASTA file containing all UCE sequence data for all taxa. We separated the FASTA file of sliced sequences by locus and aligned all loci using a parallel wrapper (seqcap_align_2.py) around MAFFT (version 7.130; Katoh *et al.* 2005). Following MAFFT alignment, we removed the locus names from all alignments, edge-and internally trimmed resulting alignments using the TRIMAL '-automated1' algorithm (Capella-Gutierrez *et al.* 2009), converted trimmed alignments back to nexus format (convert_one_align_format_to_another.py), and selected the subset of alignments (get_only_loci_with_min_taxa.py) that were 70% complete (those that contained alignment data from at least 10 of 14 taxa). We generated alignment statistics and computed the number

of informative sites across all alignments using get_align_summary_data.py and get_informative_sites.py. We concatenated the resulting alignments into a PHYLIP-formatted supermatrix (format_nexus_files_for_raxml.py), we conducted 20 maximum-likelihood (ML) searches for the phylogenetic tree that best fit the data using the unpartitioned supermatrix, RAxML (version 8.0.19; Stamatakis 2006) and the GTRGAMMA model. Following the best tree search, we used RAxML to generate 100 nonparametric bootstrap replicates, we tested bootstrap replicates for convergence, and we reconciled the best fitting ML tree with the bootstrap replicates, all using features of RAxML.

Library preparation, target enrichment and sequencing of UCEs

Following the *in silico* test of the UCE bait set, we had probes commercially synthesized as an RNA target capture array ('MYBaits'; MYcroarray, Inc.). We then extracted DNA from 30 hymenopteran species (Table S1, Supporting information) using either DNeasy extraction kits (Qiagen, Inc.) or phenol–chloroform (Maniatis *et al.* 1982) extraction procedures. We selected taxa for extraction and library preparation that span a range of divergence dates (≥220 to <5 Ma) and that represent major divisions within the order (sawflies, parasitoid wasps and stinging wasps). Following extraction we quantified DNA for each sample using a Qubit fluorometer (Life Technologies, Inc.), we randomly sheared 69–509 ng (400 ng mean) DNA to a target size of approximately 650 bp (range 400–800 bp) by sonication (Q800 or Diagenode BioRuptor; Qsonica Inc.), and we input the sheared DNA into a modified genomic DNA library preparation protocol (Kapa Biosystems) that incorporated 'with-bead' cleanup steps (Fisher *et al.* 2011) using a generic SPRI substitute (Rohland & Reich 2012; hereafter SPRI). This protocol is similar to the Kapa Biosystems protocol that uses commercial SPRI chemistry for cleanup and includes end-repair, adenylation and T/A ligation steps, except that the Fisher modification does not remove and replace SPRI beads between each step. Rather, the with-bead protocol removes and replaces a 25 mM NaCl + PEG solution, leaving the beads in-solution throughout the library preparation steps until their removal just prior to PCR amplification of the library. During adapter ligation, we also substituted custom-designed sequence-tagged adapters to the ligation reaction (Faircloth & Glenn 2012). Following adapter ligation, we PCR amplified 50% of the resulting library volume (*c.* 15 μL; 50–400 ng) using a reaction mix of 25 μL HiFi HotStart polymerase (Kapa Biosystems), 5 μL of Illumina TruSeq primer mix (5 μM each) and 5 μL double-distilled water (ddH_20) using the following thermal protocol: 98° C for 45 s; 10–12 cycles of 98° C for 15 s, 60° C for 30 s, 72° C for 60 s; and a final extension of 72° C for 5 m. We purified resulting reactions using 1X SPRI, and we rehydrated libraries in 33 μL ddH_2O. We quantified 2 μL of each library using a Qubit fluorometer. We combined groups of six libraries at equimolar ratios into enrichment pools having a final concentration of 147 ng/μL.

We prepared Cot-1 DNA from nest collections of several ant species (*Aphaenogaster fulva, Aphaenogaster rudis* and *Formica subsericea*) following the protocol of Timoshevskiy *et al.* (2012). We followed library enrichment procedures for the MYcroarray MYBaits kit (Blumenstiel *et al.* 2010), with three modifications: (i) we added 100 ng MYBaits to each reaction (a 1:5 dilution of the standard MYBaits concentration), (ii) we added 500 ng custom blocking oligos designed against our custom sequence tags and using 10 inosines to block the 10 nucleotide index sequence and (iii) for a subset of the pools (three pools, 18 samples), we tested the efficiency of our hymenopteran Cot-1 DNA by performing duplicate enrichments adding 500 ng of hymenoptera Cot-1 versus 500 ng commercially available chicken Cot-1 DNA (Applied Genetics Laboratories, Inc.). We excluded the remaining two pools from the test and used hymenoptera Cot-1 with each. We ran the hybridization reaction for 24 h at 65° C. Following hybridization, we bound all pools to streptavidin beads (MyOne C1; Life Technologies) and washed bound libraries according to a standard target enrichment protocol (Blumenstiel *et al.* 2010).

We used two different approaches for PCR recovery of the enriched libraries. For 12 of the samples (Table S1, Supporting information), we followed the standard (Blumenstiel *et al.* 2010) post-enrichment approach where we dissociated enriched DNA from RNA baits bound to streptavidin-coated beads with 0.1 N NaOH, followed by a 5-min neutralization of NaOH using an equal volume of 1 M Tris-HCl, a 1X SPRI cleanup and elution of the SPRI-purified sample in 30 μL of ddH_2O. For the remaining 18 samples, we removed the final aliquot of wash buffer following enrichment and allowed samples to dry for five minutes while sitting in a magnet stand. We removed residual buffer with sterile toothpicks. Then, we added 30 μL ddH_20 to each sample and proceeded directly to PCR recovery while the enriched libraries were still bound to streptavidin beads (Fisher *et al.* 2011). The streptavidin beads do not inhibit PCR and with-bead PCR recovery of enriched libraries is a faster and easier procedure. We combined either 15 μL of unbound, SPRI-purified, enriched library or 15 μL of streptavidin bead-bound, enriched library in water with 25 μL HiFi HotStart Taq (Kapa Biosystems), 5 μL of Illumina TruSeq primer mix (5 μM each) and 5 μL of ddH_2O.

We ran PCR recovery of each library using the following thermal profile: 98° C for 45 s; 16–18 cycles of 98° C for 15 s, 60° C for 30 s, 72° C for 60 s; and a final extension of 72° C for 5 m. We purified resulting reactions using 1.8X SPRI, and we rehydrated enriched pools in 33 μL ddH_2O. We quantified 2 μL of each enriched pool using a Qubit fluorometer.

Following quantification of the enriched pools, we verified enrichment and compared the utility of chicken Cot-1 to hymenopteran Cot-1 by designing primers (Untergasser *et al.* 2012) to amplify seven UCE loci (Table S3, Supporting information) targeted by the baits we designed. We set up a relative qPCR by amplifying two replicates of 1 ng of enriched DNA from each library at all seven loci and comparing those results to two replicates of 1 ng unenriched DNA for each library at all seven loci. We performed qPCR using a SYBR Green qPCR kit (Hoffman-LaRoche, Ltd.) on a Roche LightCycler 480. Following data collection, we computed the average of the replicate crossing point (C_p) values for each library at each amplicon for each Cot-1 treatment, and we computed fold-enrichment values, assuming an efficiency of 1.78 and using the formula 1.78^{abs}(enriched C_p − unenriched C_p).

Following qPCR verification and selection of the library pools that showed the greatest fold enrichment for a given Cot-1 treatment (chicken or hymenopteran), we diluted each pool to 2.5 ng/μL for qPCR library quantification. Using the diluted DNA, we qPCR quantified libraries using a library quantification kit (Kapa Biosystems) and assuming an average library fragment length of 500 bp. Based on the size-adjusted concentrations estimated by qPCR, we created two different equimolar pools of libraries at 10 nM concentration (Table S1, Supporting information), and we sequenced 9–10 pmol of each pool-of-pooled libraries using two runs of paired-end, 250 bp sequencing on an Illumina MiSeq (v2; UCLA Genotyping Core Facility).

Analysis of captured UCE data

We trimmed and demultiplexed FASTQ data output by BaseSpace for adapter contamination and low-quality bases using a parallel wrapper (https://github.com/faircloth-lab/illumiprocessor) around Trimmomatic (Bolger *et al.* 2014). Following read trimming, we computed summary statistics on the data using get_fastq_stats.py from the PHYLUCE package. To assemble the cleaned reads, we generated separate data sets using wrappers around the programs Trinity (version trinityrnaseq-r2013-02-25; assemblo_trinity.py; Marcais & Kingsford 2011; Grabherr *et al.* 2011) and ABySS (version 1.3.6; assemblo_abyss.py; Simpson *et al.* 2009). For both assemblies we computed coverage across assembled contigs using a program (get_trinity_coverage.py) that realigns the trimmed sequence reads to each set of assembled contigs using BWA-MEM (Li 2013), cleans the resulting BAM files using PICARD (version 1.99; http://picard.sourceforge.net/), adds read-group (RG) information to each library using PICARD, indexes the resulting BAM file using SAMTOOLS (Li *et al.* 2009) and calculates coverage at each base of each assembled contig using GATK (version 2.7.2; Van der Auwera *et al.* 2002; McKenna *et al.* 2010; DePristo *et al.* 2011).

To identify assembled contigs representing enriched UCE loci from each species, we aligned species-specific contig assemblies from both sequence assembly programs to a FASTA file of all enrichment baits using match_contigs_to_loci.py, as described above. We created a file containing the names of 30 enriched taxa from which we collected data (Table S1, Supporting information), as well as the names of 14 genome-enabled, hymenopteran taxa (Table S2, Supporting information), and we input this list to an additional program (get_match_counts.py) that queries the relational database created by matching baits to assembled contigs, as well as the relational database containing UCE match data for genome-enabled taxa (created as part of the *in silico* tests), to generate a list of UCE loci shared among all taxa. We input the list of loci generated by this program to an additional program (get_fastas_from_match_counts.py) to create a monolithic FASTA file containing all UCE sequence data for all taxa. We aligned all data in the monolithic FASTA file using MAFFT (Katoh *et al.* 2005) and seqcap_align_2.py, as described above. Following MAFFT alignment, we removed the locus names from all alignments (remove_locus_name_from_nexus_lines.py), edge- and internally trimmed resulting alignments using the '-automated1' algorithm implemented in TRIMAL (Capella-Gutierrez *et al.* 2009), converted trimmed alignments back to nexus format (convert_one_align_format_to_another.py) and selected the subset of alignments (get_only_loci_with_min_taxa.py) that were 75% complete (those that contained alignment data from at least 33 of 44 individuals). We generated alignment statistics and computed the number of informative sites across all alignments using get_align_summary_data.py and get_informative_sites.py.

We concatenated the resulting alignments into a PHYLIP-formatted supermatrix (format_nexus_files_for_raxml.py) and conducted 20 maximum-likelihood (ML) searches for the phylogenetic tree that best fit the data using the unpartitioned supermatrix, RAxML (Stamatakis 2006), and the GTRGAMMA model. Following the best tree search, we used RAxML to generate 100 nonparametric bootstrap replicates, we tested bootstrap replicates for convergence, and we reconciled the best fitting ML tree with the bootstrap replicates.

Results

Identification of UCEs

We identified 1510 nonduplicate, 60 bp regions of 100% conservation across the alignments of *apiMel4* to *nasVit2*, and we designed a capture bait set containing 2749 probes targeting these 1510 loci.

In silico test of UCEs

During our *in silico* tests, we located an average of 863.7 (95 CI: 98.3) unique UCE loci across genome-enabled hymenopteran species (Table S2, Supporting information). Following identification and filtering for uniqueness, sequence slicing, sequence alignment and trimming of resulting alignments, we generated a 70% complete matrix containing 721 UCE loci and having a mean alignment length of 1434 base pairs (95 CI: 35.5). These loci contained an average of 819 informative sites per locus, and concatenation of all loci in the complete matrix produced a supermatrix of 1 033 906 bp containing 591 033 informative sites. The phylogeny inferred from these results (Fig. S1, Supporting information) reconstructs the established relationships among genome-enabled hymenopteran lineages (Brady *et al.* 2006; Werren *et al.* 2010; Heraty *et al.* 2011; Oxley *et al.* 2014) with complete support.

In vitro test of UCEs

We extracted an average of 1894 ng DNA (181–6480 ng) from each hymenopteran species and input an average of 400 ng (69–509 ng) to the library preparation process. Following library prep, PCR amplification and SPRI purification, DNA libraries contained approximately 100 ng DNA (53–151 ng). Fold-enrichment values of enriched libraries estimated by qPCR suggested that commercial chicken Cot-1 performed better than the hymenopteran Cot-1 we prepared by approximately 500-fold (Table S4, Supporting information): pooled libraries blocked with chicken Cot-1 showed an average fold enrichment of 744x while pooled libraries blocked with hymenopteran Cot-1 showed an average fold enrichment of 178x. Based on these results, we sequenced the three enriched pools where we could choose chicken Cot-1 as blocking DNA, as well as the remaining two pools where we could only choose hymenopteran Cot-1 as blocking DNA.

Sequencing produced an average of 1.1 million (95 CI: 249, 342) reads per sample. Reads averaged 192.6 bp (95 CI: 3.7) following demultiplexing, quality- and adapter-trimming (Table S5, Supporting information). Using Trinity (Table 1), we assembled these DNA reads into an average of 74 140 contigs of 347.7 bp in length (95 CI: 6.9) and having a mean coverage of 4.1X (95 CI: 0.3). Using ABySS (Table S6, Supporting information), we assembled these DNA reads into an average of 143 863 contigs of 202.7 bp in length (95 CI: 5.1) and having a mean coverage of 3.8X (95 CI: 0.2).

After searching for UCEs within the Trinity assemblies (Table 1; Table S7, Supporting information), we enriched an average of 721 (95 CI: 48.2) unique UCE loci, the average locus length was 1010 bp (95 CI: 66.1), the average coverage per enriched UCE locus was 52.3X (95 CI: 9.4), and the mean percentage of reads-on-target was 30% (95 CI: 2.7%). When searching against the ABySS assemblies (Tables S6 and S8, Supporting information), we enriched an average of 477 (95 CI: 56.4) unique UCE loci, the average locus length was 669.1 bp (95 CI: 36.9), the average coverage per enriched UCE locus was 40.7X (95 CI: 5.1), and the mean percentage of reads-on-target was 12.5% (95 CI: 3%). These and other summary statistics on assemblies (Fig. S2, Supporting information) suggest that Trinity-assembled UCE contigs have more desirable properties for downstream phylogenetic analyses, in aggregate, than ABySS assemblies.

Following alignment of the Trinity-assembled data, alignment trimming and filtering of loci having fewer than 33 taxa (75% complete), we retained 600 alignments having an average length of 691.4 bp (95 CI: 44.4 bp). The average number of taxa present in these 600 alignments was 39.2 (95 CI: 0.2). The concatenated, Trinity supermatrix contained 414 849 bp, 413 782 total nucleotide characters and 282 973 informative sites. Following alignment of the ABySS -assembled data, alignment trimming and filtering of loci having fewer than 33 taxa (75% complete), we retained 196 alignments of 522.5 bp (95 CI: 82.1 bp) in length. The average number of taxa present in these 196 alignments was 36.71 (95 CI: 0.37). The concatenated, ABySS supermatrix contained 102 418 bp, 102 148 total nucleotide characters and 60 714 informative sites.

We inferred a phylogeny from both Trinity assemblies (Fig. 1) and ABySS assemblies (Fig. S3, Supporting information). Because the Trinity assemblies produced a larger number of longer, higher coverage UCE loci that yielded a larger, 70% complete, concatenated supermatrix, we focus on the relationships we inferred from the Trinity data. However, the ABySS topology (Fig. S3, Supporting information), while having slightly lower support at several nodes, was identical to the topology we inferred from the Trinity assemblies.

Generally, the relationships among Hymenoptera we inferred from the Trinity supermatrix (Fig. 1) accurately reconstructed: (i) the relationships among genome-enabled hymenopterans inferred during our *in silico* analysis, (ii) the established relationships between taxa

Table 1 Summary values describing the number of contigs assembled by Trinity from adapter- and quality-trimmed reads ('All' contigs), their average coverage, the mean length of All contigs, the count of unique reads aligned to All contigs, the number of ultraconserved element (UCE) contigs identified from the pool of All contigs, the mean length of UCE contigs, the average UCE contig sequencing coverage and the percentage of unique reads that aligned to UCE contigs (this is a percentage of the percentage of unique reads aligning to All contigs)

Taxon	All contigs	All contigs coverage	All contigs coverage 95 CI	All contigs mean length	All contigs mean length 95 CI	All contigs unique reads aligned	UCE contigs	UCE contigs mean length	UCE contigs coverage	UCE contigs unique reads aligned
Acordulecera pellucida	30 033	3.4	0.2	344.9	2.6	70.7%	341	1025.0	26.3	18.4%
Andrena (Callandrena) asteris	4587	4.6	0.3	358.0	6.2	69.2%	740	574.7	9.8	44.4%
Andrena (Melandrena) sp	33 761	3.3	0.3	345.1	3.4	68.5%	774	857.0	18.2	25.9%
Aphaenogaster albisetosa	157 813	3.9	0.1	359.4	1.0	78.7%	764	1128.6	88.3	26.0%
Aphaenogaster megommata	117 378	4.5	0.3	341.6	1.2	76.6%	751	1184.1	79.2	28.8%
Aphaenogaster tennesseensis	76 656	4.8	0.1	336.8	1.7	67.1%	751	1056.3	62.4	30.5%
Aphaenogaster texana	49 221	4.9	0.1	330.3	2.3	68.9%	750	924.6	51.5	33.2%
Aporus niger	16 552	6.3	2.7	332.5	3.8	68.2%	740	714.1	14.5	17.7%
Bombus pensylvanicus	26 877	3.2	0.1	321.4	2.7	72.2%	780	859.7	21.4	35.0%
Chalybion californicus	44 727	3.8	0.6	324.1	1.4	69.0%	778	812.2	33.2	29.6%
Chyphotes mellipes	105 477	4.4	0.7	383.5	1.6	79.3%	774	1184.0	66.2	26.6%
Evaniella semaeoda	18 980	4.5	0.2	359.1	4.3	73.6%	638	971.7	31.1	39.1%
Messor piceus	91 858	4.6	0.8	332.6	1.4	71.7%	730	1111.7	58.9	26.2%
Metapolybia cingulata	63 299	3.3	0.1	326.0	1.3	77.0%	685	823.3	40.1	24.7%
Mischocyttarus flavitarsis	16 624	5.5	1.6	330.2	3.6	82.1%	634	711.2	30.0	32.4%
Nasonia vitripennis	27 195	4.9	0.2	314.2	2.1	77.2%	1166	771.1	46.9	57.1%
Nematus tibialis	48 874	3.5	0.1	350.3	2.1	72.4%	453	1049.5	47.9	26.4%
Orthogonalys pulchella	106 246	4.1	0.1	405.2	1.4	87.9%	706	1364.0	109.0	35.0%
Pogonomyrmex occidentalis	154 514	3.9	0.1	362.2	1.0	83.5%	741	1142.4	97.5	26.8%
Sapyga pumila	108 990	4.0	0.1	361.6	1.4	77.5%	720	1046.9	86.4	28.6%
Scolia verticalis	55 545	3.9	0.3	350.7	1.9	75.5%	760	1070.4	56.6	36.0%
Sericomyrmex harekulli	25 698	3.5	0.1	329.0	2.6	71.3%	744	814.8	22.3	33.5%
Stenamma diecki	108 642	3.9	0.1	365.8	1.7	71.8%	751	1142.0	53.5	23.7%
Stenamma expolitum	135 131	3.7	0.1	363.2	1.3	76.9%	749	1212.1	69.3	25.7%
Stenamma felixi	138 761	3.8	0.1	341.7	1.1	77.7%	762	1071.8	75.3	25.1%
Stenamma impar	89 581	4.4	0.3	355.3	2.1	68.7%	741	1056.0	49.8	22.4%
Stenamma megamanni	78 363	3.0	0.0	354.6	1.6	75.8%	754	1138.0	37.8	28.6%
Stenamma megamanni2	147 772	3.8	0.1	359.5	1.3	79.7%	756	1232.9	87.5	28.7%
Stenamma muralla	102 541	3.4	0.1	334.9	1.1	79.7%	734	1132.0	61.6	30.6%
Taxonus pallidicornis	42 507	3.3	0.1	356.4	2.5	71.6%	459	1140.9	37.7	27.5%

from which we collected data, *de novo* and (iii) the established relationships between genome-enabled taxa and species from which we collected data (Danforth *et al.* 2006; Brady *et al.* 2006; Werren *et al.* 2010; Heraty *et al.* 2011; Branstetter 2012; Oxley *et al.* 2014; Ward *et al.* 2014).

In Fig. 1, the sawflies, represented here by only the superfamily Tenthredinoidea, formed a clade sister to the Apocrita. Within the Apocrita, parasitic wasps formed a paraphyletic grade leading to a monophyletic Aculeata (stinging wasps, ants and bees) with *Orthogonalys* (Trigonalidae)+*Evaniella* (Evaniidae) recovered as sister to the aculeates. Within Aculeata, we recovered five main groups with maximum support (note that we did not include chrysidoid wasps): ants (Formicidae), spheciform bees+wasps (Apoidea), vespid wasps (Vespidae), scoliid wasps (Scoliidae) and tiphioid-pompiloid wasps (Chyphotidae+Pompilidae+Sapygidae). Among these groups, we inferred the ants to be sister to a clade containing all remaining aculeate lineages with maximum support. Within the clade containing the remaining aculeates, we recovered the Scoliidae as sister to the Apoidea (87% support), and we recovered the Vespidae as sister to the tiphioid-pompiloid wasps (58% support). Within the ants, we recovered all expected relationships among the five included subfamilies (Ponerinae, Dorylinae, Dolichoderinae, Formicinae and Myrmicinae; Brady *et al.* 2006; Moreau & Bell 2013), and several closely related ant genera and species belonging to the tribe Stenammini (*Aphaenogaster*, *Messor* and *Stenamma*; Branstetter 2012; Ward *et al.* 2014) with high (≥87%) support.

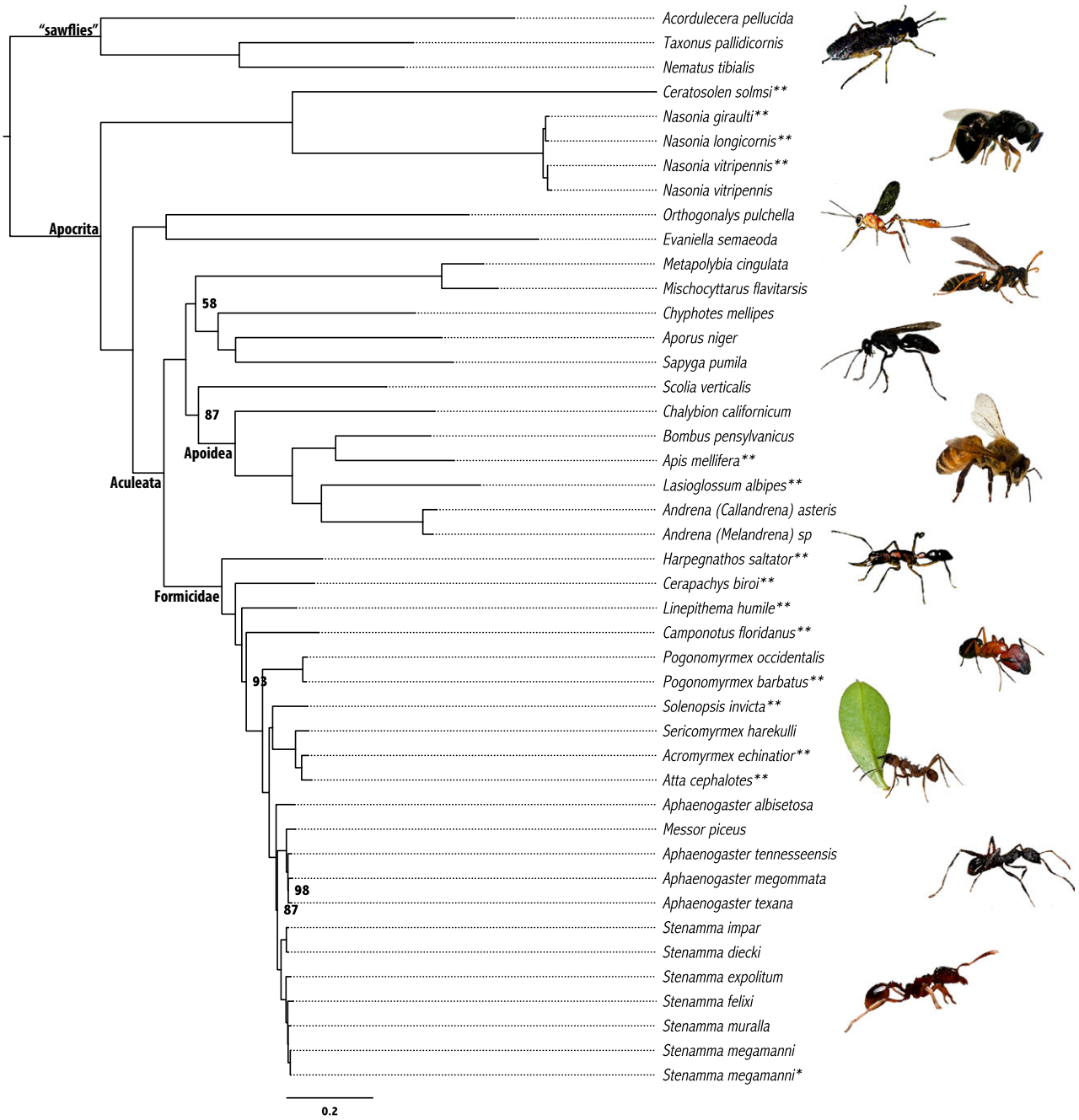

Fig. 1 Maximum-likelihood phylogeny inferred from a 75% complete supermatrix containing data from 14 genome-enabled taxa (identified by double-asterisks) and 30 taxa from which we enriched and assembled (Trinity) ultraconserved element loci. We show bootstrap support values only where support is <100%, and the single asterisk beside *Stenamma megamanni* denotes that this sample represents a different population of the same species.

To test the effects of removing distantly related sawfly lineages on the topology and support inferred across the UCE data, we constructed a new UCE data set lacking sawfly lineages because the sawfly data were the most incomplete, with respect to counts of recovered loci across all taxa (see Fig. S4, Supporting information and below), and the inclusion of sawflies had the largest effect on the size of our incomplete matrix. This new data set (75% complete) included 638 UCE loci, contained an average of 37.2 taxa (95 CI: 0.2), and had an average alignment length of 737.1 bp (95 CI: 46.4). The supermatrix contained 470 258 bp, 469 081 total nucleotide characters and 310 253 (+27 280) informative sites. Following inference from this updated data set with RAxML using

approaches identical to those described above, the resulting phylogeny (Fig. S6, Supporting information) had the same topology as the tree including sawflies with the exception of inferred relationships between two nonaculeate taxa, *Evaniella* and *Orthognalys*.

Analysis of capture success

Based on the differences in capture success we observed across the resulting phylogeny (Fig. S4, Supporting information), we analysed several summary metrics (Tables S7 and S8, Supporting information), *post hoc,* using general linear models (R, version 2.5.12; Team 2011) to investigate those parameters affecting the number (Poisson link function) and length (Gaussian link function) of UCE loci we recovered. With these values, we also included an explicit measure of pairwise genetic distance between all taxa from which we enriched sequence data and the *nasVit2* genomic assembly, from which we designed capture baits. We estimated distance values from the concatenated Trinity supermatrix using the 'distance' method of PyCogent (version 1.5.3; Knight *et al.* 2007) and assuming a GTR site rate substitution model. We used Akaike's information criteria (AIC) to rank and compare linear models, and we model-averaged estimates across parameters where there was a valid set ($w_i > 0.10$; Royall 1997) of candidate models. These *post hoc* analyses suggest that UCE capture success may be driven by several factors, in addition to phylogenetic distance between the probe design source and the taxa being enriched. Specifically, Akaike weights suggest that a 'global' model containing four parameters (distance + reads + mean read length + assembly method) best approximates the data (Table S9, Supporting information), that there are large differences among parameter effect sizes (Fig. S5, Supporting information), and that phylogenetic distance has the largest effect of parameters we investigated on the number of UCE contigs enriched. The size of this effect is tempered somewhat when considering only the Trinity assemblies, where read length appears to play a role (Table S10, Fig. S5, Supporting information). Similarly, length of enriched UCE contigs may best be explained (Table S11, Supporting information) by a global model containing three parameters (distance + reads + assembly method), assembly method probably plays a larger role in resulting length of UCE loci, and phylogenetic distance retains a large effect on resulting contig length (Fig. S5, Supporting information). When considering only the Trinity assemblies, the effects of distance and the number of reads are both important factors affecting resulting contig length (Fig. S12, Supporting information). In all of these results, it is important to keep in mind that phylogenetic distance falls on the interval [0,1], so the effect size of this parameter is tempered by typically small changes in its value.

Discussion

We have developed a powerful new genomic tool for estimating phylogenetic relationships among members of the hyperdiverse insect order Hymenoptera. By extending and improving prior work (Faircloth *et al.* 2012), we identified over 1500 highly conserved genomic regions between distantly related Hymenoptera taxa, collected these loci from 14 genome-enabled and 30 non-genome-enabled taxa using *in silico* and *in vitro* techniques and used the resulting genome-scale sequence data to accurately infer both deep (*c.* 220–300 Ma) and relatively shallow (≤1 Ma) relationships. Although other phylogenomic approaches have been employed among arthropods (Johnson *et al.* 2013), this is the first time that sequence capture of conserved regions has been used to collect genome-scale DNA data from this group.

Compared to recent phylogenetic studies investigating higher-level relationships within Hymenoptera (Sharkey 2007; Heraty *et al.* 2011; Klopfstein *et al.* 2013), the UCE data recovered all well-established relationships with complete support. In addition, the UCE data suggest a novel relationship within the Aculeata, in which the ants are sister to all remaining aculeate lineages included here. The aculeates contain all major lineages of social insects (except termites) including ants, vespid wasps and several lineages of social bees. Aculeata also includes the most important group of pollinators (bees). Hence, understanding relationships among the aculeates is critical to provide the comparative framework needed to study the origins and evolution of sociality and pollination biology in this group (Danforth 2013). Until recently, phylogenetic studies of aculeates have been based on a relatively small number of characters and have produced conflicting results (Brothers 1999; Pilgrim *et al.* 2008; Peters *et al.* 2011; Debevec *et al.* 2012). A recent transcriptome-based study (Johnson *et al.* 2013) sequenced key lineages within Aculeata and produced a fully resolved phylogeny of aculeate lineages, recovering a novel relationship in which ants are sister to the Apoidea (spheciform bees+wasps). Our UCE data set did not recover this relationship. Instead, we found ants to be sister to all remaining aculeate lineages with complete support, but there were several nodes within each clade receiving moderate (≥58%) support. Our study also differed from Johnson *et al.* (2013) in the placement of vespid wasps as sister to the tiphioid-pompiloid wasps (Chyphotidae+Pompilidae+Sapygidae) and the scoliid wasps as sister to the spheciform wasps+bees (Apoidea). Previous work by Debevec *et al.* (2012) also recovered

this placement of scoliid wasps as sister to the spheciform wasps+bees.

Given the importance of resolving relationships among aculeate lineages, we tested the effects of removing sawfly lineages on the topology and support inferred across the UCE tree presented in Fig. 1. Following inference from this updated data set with RA*x*ML, the resulting phylogeny (Fig. S6, Supporting information) had the same topology as the tree including sawflies, except that in Fig. 1, two nonaculeate taxa, *Evaniella* and *Orthognalys* form a clade with maximum support, while in Fig. S6 (Supporting information), these taxa form a grade, also with maximum support. Support values for internal nodes were marginally higher in the tree excluding sawflies. The stability of the recovered relationships within Aculeata between these two trees and across different assembly methods suggests that neither the count of loci, nor the total amount of data, nor the assembly approach are driving the differences we observed between our results and those of Johnson *et al.* (2013).

Rather, taxon sampling (e.g. our study does not include any chrysidoid wasps) or other differences among each data set including size, analytical approach, nucleotide composition, locus type, the number of independent loci sampled and matrix completeness could explain the differences in topology we observed. For example, Johnson *et al.* (2013) collected and analysed both larger and smaller amounts of data (175 404–3 001 657 sites) of a different type (amino acid residues) from fewer taxa (n = 19) that included variable counts of loci (308–5214 genes) spanning a range of matrix completeness (50–100%), and they inferred their phylogeny using concatenated maximum likelihood, concatenated Bayesian and summary-statistic gene tree species tree approaches. In contrast, we collected and analysed a less variable amount of data (102 418–469 081 sites), from a larger number of taxa (n = 41–43) that included variable counts of loci (196 – 638 loci) spanning a small range of matrix completeness (70–75%). We inferred the phylogeny using a concatenated maximum-likelihood approach. The types of differences between these two studies and their effects on phylogenetic reconstruction are the sorts of questions that deserve the bulk of current and future analytical effort in phylogenomics.

Focusing within ants, we captured an average of 748 UCE loci (95 CI: 5.0) using the bait set we designed and inferred nearly all relationships with complete support. The relationships we recovered among ant subfamilies agree with several recent molecular phylogenies of ants (Moreau *et al.* 2006; Brady *et al.* 2006; Moreau & Bell 2013). Furthermore, most relationships within the tribe Stenammini (*Aphaenogaster*, *Messor* and *Stenamma*), including relationships within *Stenamma*, agree with recent molecular studies (Moreau *et al.* 2006; Brady *et al.* 2006; Branstetter 2012; Moreau & Bell 2013). Our study also agrees with a recent 11-gene phylogeny that documents the nonmonophyly of the genus *Aphaenogaster* (Ward *et al.* 2014). These observations are important because they demonstrate the potential for using UCEs to resolve shallow relationships within the Hymenoptera (divergence dates among *Stenamma* species are estimated at *c.* 35 to <5 Ma; Branstetter 2012) similar to results from UCE data collected among vertebrates (Faircloth *et al.* 2013; Smith *et al.* 2014).

A major advantage of the UCE approach we describe over transcriptome-based methods is that it does not require specially preserved tissues. Here, we successfully extracted and enriched DNA from insect specimens that ranged from 12 years old to weeks old using a variety of collection methods, including several that were suboptimal for DNA preservation (ethanol preserved or dry pinned) and resulted in the extraction of little DNA (Table S1, Supporting information). Furthermore, we successfully generated and enriched UCE loci from genomic libraries constructed using as little as 70 ng of DNA. This finding is significant because many arthropod taxa are small, yielding very low amounts of DNA, and our results suggest we can successfully prepare and enrich libraries from low DNA inputs. New library preparation approaches, including the Hyper Prep Kit (Kapa Biosystems) and the NEBNext Ultra Kit (New England Biolabs), should make it possible to use even less DNA in the future without resorting to expensive modifications of protocol. The ability to use small, moderately old and sometimes low-quality specimens with the UCE approach we describe means that much of the available materials in museums and other collections can be used as a DNA source for phylogenomic studies – making it possible to sequence very rare and, often, very important taxa.

Acknowledgements

We thank Alex Buerkle, Jennifer Mandel and three anonymous reviewers for their comments, which improved this manuscript. We thank Dave Smith, Ana Jesovnik, Jack Longino, Phil Ward, Brian Fisher, Peter Hawkes, Jim Carpenter, James Pitts, Gary Hevel and Sam Droege for providing some of the tissue samples we used. Hong Zhao helped with DNA extraction, and Alex Chase assisted with library preparation. We thank Uma Dandekar and Hemani Wijesuriya for performing MiSeq runs and Travis Glenn for helpful discussions. Phil Ward provided insightful comments regarding our results and relationships within the Aculeata. To make the cartoon Hymenoptera used in Fig. 1, we altered open-access or public domain photographs taken by the following individuals: Bob Peterson (*Camponotus* sp.), Srikaanth Sekar (*Harpegnathos* sp.), Gilles San Martin (*Stenamma* sp.), Christophe Quintin (*Taxonus* sp.), Franco Folini (*Apis* sp.), M.E. Clark (*Nasonia* sp.), Muhammad Mahdi Karim (Cremastinae), OpenCage (Vespidae) and DeadStar (*Acromyrmex* sp.). All

modified images are available as PSD files from figshare (doi:10.6084/m9.figshare.1173286).

Funding

A Smithsonian Institution grant from the Consortium for Understanding and Sustaining a Biodiverse Planet (to SGB, BCF, and NDW) provided most of the funding for this project and support for NDW's participation. A Smithsonian Institution Buck Postdoctoral Fellowship and NSF grants DEB-1354739 and DEB-1354996 (project ADMAC) supported MGB. NSF DEB-1242260 (to BCF) and an Amazon Web Services Education Grant (to BCF) supported computational portions of this work. NSF EF-0431330 (to SGB) provided resources to obtain some specimens used in this study. This study was also supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

References

Ahituv N, Zhu Y, Visel A *et al.* (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biology*, **5**, e234.

Bejerano F, Pheasant M, Makunin I *et al.* (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.

Beye M, Moritz RF, Epplen C (1994) Sex linkage in the honeybee *Apis mellifera* detected by multilocus DNA fingerprinting. *Die Naturwissenschaften*, **81**, 460–462.

Blumenstiel B, Cibulskis K, Fisher S *et al.* (2010) Targeted exon sequencing by in-solution hybrid selection. *Current Protocols in Human Genetics*, **Chapter 18**, Unit 18.4.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bourke AFG, Franks NR (1995) *Social Evolution in Ants*. Princeton University Press, Princeton, New Jersey.

Bradley TJ, Briscoe AD, Brady SG *et al.* (2009) Episodes in insect evolution. *Integrative and Comparative Biology*, **49**, 590–606.

Brady SG, Schultz TR, Fisher BL, Ward PS (2006) Evaluating alternative hypotheses for the early evolution and diversification of ants. *Proceedings of the National Academy of Sciences, USA*, **103**, 18172–18177.

Branstetter MG (2012) Origin and diversification of the cryptic ant genus Stenamma Westwood (Hymenoptera: Formicidae), inferred from multilocus molecular data, biogeography and natural history. *Systematic Entomology*, **37**, 478–496.

Brothers DJ (1999) Phylogeny and evolution of wasps, ants and bees (Hymenoptera, Chrysidoidea, Vespoidea and Apoidea). *Zoologica Scripta*, **28**, 233–250.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Crawford NG, Faircloth BC, McCormack JE *et al.* (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, **8**, 783–786.

Dalling JW, Swaine MD, Garwood NC (1998) Dispersal patterns and seed bank dynamics of pioneer trees in moist tropical forest. *Ecology*, **79**, 564–578.

Danforth BN (2013) Social Insects: Are ants just wingless bees? *Current Biology*, **23**, R1011–R1012.

Danforth BN, Sipes S, Fang J, Brady SG (2006) The history of early bee diversification based on five genes plus morphology. *Proceedings of the National Academy of Sciences, USA*, **103**, 15118–15123.

Danforth BN, Cardinal S, Praz C, Almeida EAB, Michez D (2013) The impact of molecular data on our understanding of bee phylogeny and evolution. *Annual Review of Entomology*, **58**, 57–78.

Davis RB, Baldauf SL, Mayhew PJ (2010) The origins of species richness in the Hymenoptera: insights from a family-level supertree. *BMC evolutionary biology*, **10**, 109.

Debevec AH, Cardinal S, Danforth BN (2012) Identifying the sister group to the bees: a molecular phylogeny of Aculeata with an emphasis on the superfamily Apoidea. *Zoologica Scripta*, **41**, 527–535.

Del Toro I, Ribbons RR, Pelini SL (2012) The little things that run the world revisited: a review of ant-mediated ecosystem services and disservices (Hymenoptera: Formicidae). *Myrmecological News*, **17**, 133–146.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Derti A, Roth FP, Church GM, Wu C-T (2006) Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nature Genetics*, **38**, 1216–1220.

Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One*, **7**, e42543.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717–726.

Faircloth BC, Sorenson L, Santini F, Alfaro ME (2013) A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *PLoS One*, **8**, e65923.

Fisher S, Barry A, Abreu J *et al.* (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology*, **12**, R1.

Fittkau EJ, Klinge H (1973) On biomass and trophic structure of the central Amazonian rain forest ecosystem. *Biotropica*, **5**, 2–14.

Gaston KJ (1991) The magnitude of global insect species richness. *Conservation Genetics*, **5**, 283–296.

Gaston KJ, Gauld ID, Hanson P (1996) The size and composition of the hymenopteran fauna of Costa Rica. *Journal of Biogeography*, **23**, 105–113.

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

Grimaldi D, Engel MS (2005) *Evolution of the Insects*. Cambridge University Press, Cambridge, UK.

Harmston N, Baresic A, Lenhard B (2013) The mystery of extreme non-coding conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**, 20130021.

Harris RS (2007) *Improved Pairwise Alignment of Genomic DNA*. The Pennsylvania State University, University Park, PA.

Heraty J, Ronquist F, Carpenter JM *et al.* (2011) Evolution of the hymenopteran megaradiation. *Molecular phylogenetics and evolution*, **60**, 73–88.

Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, **443**, 931–949.

Howard KJ, Thorne BL (2010) Eusocial evolution in termites and Hymenoptera. In: *Biology of Termites: A Modern Synthesis* (eds Bignell DE, Roisin Y, Lo N), pp. 97–132. Springer, Dordrecht.

Hunt GJ, Page RE (1994) Linkage analysis of sex determination in the honey-bee (*Apis mellifera*). *Molecular & General Genetics*, **244**, 512–518.

Johnson BR, Linksaver TA (2010) Deconstructing the superorganism: Social physiology, groundplans, and sociogenomics. *Quarterly Review of Biology*, **85**, 57–79.

Johnson BR, Borowiec ML, Chiu JC *et al.* (2013) Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. *Current Biology*, **23**, 2058–2062.

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511–518.

Kevan PG (1999) Pollinators as bioindicators of the state of the environment: species, activity and diversity. *Agriculture, Ecosystems & Environment*, **74**, 373–393.

Klopfstein S, Vilhelmsen L, Heraty JM, Sharkey M, Ronquist F (2013) The hymenopteran tree of life: evidence from protein-coding genes and objectively aligned ribosomal data (eds Klopfstein S, Vilhelmsen L, Heraty JM, Sharkey M, Ronquist F). *PLoS One*, **8**, e69344.

Knight R, Maxwell P, Birmingham A *et al.* (2007) PyCogent: a toolkit for making sense from sequence. *Genome Biology*, **8**, R171.

Kremen C, Williams NM, Thorp RW (2002) Crop pollination from native bees at risk from agricultural intensification. *Proceedings of the National Academy of Sciences, USA*, **99**, 16812–16816.

LaSalle J, Gauld ID (1993) Hymenoptera: their biodiversity, and their impact on the diversity of other organisms. In: *Hymenoptera and Biodiversity* (eds LaSalle J, Gauld ID), pp. 1–26. CAB International, Wallingford, UK.

Levey DJ, Byrne MM (1993) Complex ant-plant interactions: rain-forest ants as secondary dispersers and post-dispersal seed predators. *Ecology*, **74**, 1802.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. available from: arXiv.org.

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

Maniatis T, Fritsch E, Sambrook J (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Marcais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

McCormack JE, Faircloth BC, Crawford NG *et al.* (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species tree analysis. *Genome research*, **22**, 746–754.

McCormack JE, Harvey MG, Faircloth BC *et al.* (2013) A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS One*, **8**, e54848.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–1303.

Mehdiabadi NJ, Mueller UG, Brady SG *et al.* (2012) Symbiont fidelity and the origin of species in fungus-growing ants. *Nature Communications*, **3**, 840.

Michener CD (2007) *The Bees of the World*. Johns Hopkins University Press, Baltimore, Maryland.

Moreau CS (2009) Inferring ant evolution in the age of molecular data (Hymenoptera: Formicidae). *Myrmecological News*, **12**, 201–210.

Moreau CS, Bell CD (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 CS, Bell CD, Vila R, Archibald SB, Pierce NE (2006) Phylogeny of the ants: diversification in the age of angiosperms. *Science*, **312**, 101–104.

Mueller UG, Rabeling C (2008) A breakthrough innovation in animal evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 5287–5288.

Mueller UG, Gerardo NM, Aanen DK, Six DL, Schultz TR (2005) The evolution of agriculture in insects. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 563–595.

Oxley PR, Ji L, Fetter-Pruneda I *et al.* (2014) The genome of the clonal raider ant *Cerapachys biroi*. *Current Biology*, **24**, 451–458.

Page RE, Gadau J, Beye M (2002) The emergence of hymenopteran genetics. *Genetics*, **160**, 375–379.

Peters RS, Meyer B, Krogmann L *et al.* (2011) The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biology*, **9**, 55.

Pilgrim EM, von Dohlen CD, Pitts JP (2008) Molecular phylogenetics of Vespoidea indicate paraphyly of the superfamily and novel relationships of its component families and subfamilies. *Zoologica Scripta*, **37**, 539–560.

Quicke DL (1997) *Parasitic Wasps*. Chapman & Hall, London.

Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.

Ronquist F, Klopfstein S, Vilhelmsen L *et al.* (2012) A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology*, **61**, 973–999.

Roubik DW (1989) *Ecology and Natural History of Tropical Bees*. Cambridge University Press, Cambridge.

Roubik DW (1995) *Pollination of Cultivated Plants in the Tropics*. Food and Agriculture Organization of the United Nations, Rome.

Royall R (1997) *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London, UK.

Sandelin A, Bailey P, Bruce S *et al.* (2004) Arrays of ultraconserved noncoding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, **5**, 99.

Savolainen R, Vepsäläinen K (2003) Sympatric speciation through intraspecific social parasitism. *Proceedings of the National Academy of Sciences, USA*, **100**, 7169–7174.

Schultz TR, Brady SG (2008) Major evolutionary transitions in ant agriculture. *Proceedings of the National Academy of Sciences, USA*, **105**, 5435–5440.

Sharanowski BJ, Robbertse B, Walker J *et al.* (2010) Expressed sequence tags reveal Proctotrupomorpha (minus Chalcidoidea) as sister to Aculeata (Hymenoptera: Insecta). *Molecular phylogenetics and evolution*, **57**, 101–112.

Sharkey MJ (2007) Phylogeny and classification of Hymenoptera. *Zootaxa*, **1668**, 521–548.

Siepel A, Bejerano G, Pedersen JS *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, **15**, 1034–1050.

Simpson J, Wong K, Jackman S *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome research*, **19**, 1117–1123.

Smit AFA, Hubley R, Green P (1996–2010) *RepeatMasker Open-3.0*. http://www.repeatmasker.org.

Smith C, Toth A, Suarez A, Robinson G (2008) Genetic and genomic analyses of the division of labour in insect societies. *Nature Reviews Genetics*, **9**, 735–748.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT (2014) Target capture and massively parallel sequencing of ultraconserved elements (UCEs) for comparative studies at shallow evolutionary time scales. *Systematic Biology*, **63**, 83–95.

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

Team R (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Tewhey R, Nakano M, Wang X *et al.* (2009) Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biology*, **10**, R116.

Timoshevskiy VA, Sharma A, Sharakhov IV, Sharakhova MV (2012) Fluorescent *in situ* hybridization on mitotic chromosomes of mosquitoes. *Journal of Visualized Experiments*, **67**, e4215.

Untergasser A, Cutcutache I, Koressaar T *et al.* (2012) Primer3-new capabilities and interfaces. *Nucleic Acids Research*, **40**, e115.

Van der Auwera GA, Carneiro MO, Hartl C *et al.* (2002) *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*. John Wiley & Sons Inc, Hoboken, New Jersey.

Wang J, Wurm Y, Nipitwattanaphon M *et al.* (2013) A Y-like social chromosome causes alternative colony organization in fire ants. *Nature*, **493**, 664–668.

Ward PS, Brady SG, Fisher BL, Schultz TR (2014) The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae). *Systematic Entomology*, doi: 10.1111/syen.12090.

Werren JH, Richards S, Desjardins CA *et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, **327**, 343–348.

Wilson EO, Hölldobler B (2005) The rise of the ants: a phylogenetic and ecological explanation. *Proceedings of the National Academy of Sciences, USA*, **102**, 7411–7414.

S.G.B., B.C.F. and N.D.W. contributed text to the grant application supporting this work. S.G.B., B.C.F., M.G.B. and N.D.W. designed the study. B.C.F., N.D.W. and M.G.B. conducted the laboratory work. B.C.F. analysed the data. B.C.F., M.G.B. and S.G.B. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Data accessibility

The bait set and enrichment protocols used as part of this manuscript are available under Creative Commons licene (CC-BY-3.0) from http://ultraconserved.org. The bait set is also available from Dryad (doi: 10.5061/dryad.46195). Computer programs used throughout this study are available from
https://github.com/faircloth-lab/uce-probe-design,
http://github.com/faircloth-lab/phyluce, and
https://github.com/faircloth-lab/illumiprocessor under an open-source, BSD-style licence. Sequence reads generated as part of this manuscript are available from the NCBI Sequence Read Archive (SRA PRJNA248919). Trinity contig assemblies representing UCE loci are available from GenBank (KM411995–KM433620). Additional data including the probe set design file, sequence alignments, alignment supermatrices, configuration files, estimated phylogenetic distances, qPCR results and inferred trees are also available from Dryad (doi: 10.5061/dryad.46195).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Family, species, collection identifier, collection year, collection country, collection method, voucher identifier, voucher depository, total amount of extract DNA, amount of DNA input to library preparation, post-enrichment method, and MiSeq run of all samples used for target enrichment.

Table S2 Species, genome assembly, genome assembly source, reference, and number of UCE loci in assembly for all genome-enabled taxa.

Table S3 Quantitative PCR primers used for assessment (relative quantification) of enrichment success and enrichment differences of Cot-1 sources.

Table S4 Crossing point (C_p) values for quantitative PCR showing the fold enrichment differences between unenriched controls, enrichments using chicken Cot-1 as a blocking agent, enrichments using hymenoptera Cot-1 as a blocking agent, and Δ Cot-1 or the fold-enrichment difference between chicken and hymenoptera Cot-1.

Table S5 Summary values describing the number of reads collected during sequencing of each enriched library.

Table S6 Summary values describing the number of contigs assembled by ABySS from adapter- and quality-trimmed reads ("All" contigs), their average coverage, the mean length of All contigs, the count of unique reads aligned to All contigs, the number of UCE contigs identified from the pool of All contigs, the mean length of UCE contigs, the average UCE contig sequencing coverage, and the percentage of unique reads that aligned to UCE contigs (this is a percentage of the percentage of unique reads aligning to All contigs).

Table S7 Summary values describing attributes of the UCE contigs assembled by Trinity.

Table S8 Summary values describing attributes of the UCE contigs assembled by ABySS.

Table S9 Model structure, AIC, number of parameters, AICc, and Akaike weight (w_i) for general linear models of parameters affecting the mean number of UCE contigs captured.

Table S10 Model structure, AIC, number of parameters, AICc, and Akaike weight (w_i) for general linear models of parameters affecting the number of UCE contigs captured among Trinity (only) assemblies.

Table S11 Model structure, AIC, number of parameters, AICc, and Akaike weight (w_i) for general linear models of parameters affecting the length of UCE contigs captured.

Table S12 Model structure, AIC, number of parameters, AICc, and Akaike weight (w_i) for general linear models of parameters affecting the length of UCE contigs captured among Trinity (only) assemblies.

Fig. S1 Maximum likelihood phylogeny inferred from a 75% complete supermatrix containing data from ultraconserved elements identified in 14 genome-enabled taxa.

Fig. S2 Box plots showing differences in standard metrics among UCE contigs assembled by Trinity or ABySS.

Fig. S3 Maximum likelihood phylogeny inferred from a 75% complete supermatrix containing data from 14 genome-enabled taxa (identified by double-asterisks) and 30 taxa from which we enriched and assembled (ABySS) ultraconserved element loci.

Fig. S4 The topology from Figure 1, with branches colored to indicate the approximate number of ultraconserved element loci we captured, by taxon, relative to the total number of loci captured from *Nasonia vitripennis* (n = 1166).

Fig. S5 Bar plots comparing parameter (β) estimates (±95% CI) from general linear models of factors affecting the number of UCE contigs enriched and the length of enriched UCE contigs.

Fig. S6 Maximum likelihood phylogeny inferred from a 75% complete supermatrix containing data from 14 genome-enabled taxa (identified by double-asterisks) and 27 taxa from which we enriched and assembled (Trinity) ultraconserved element loci.