

1 **Draft genome assembly and population genetics of an agricultural pollinator, the solitary**  
2 **alkali bee (Halictidae: *Nomia melanderi*)**

3 Karen M. Kapheim<sup>1,2\*</sup>, Hailin Pan<sup>3,4,5</sup>, Cai Li<sup>6</sup>, Charles Blatti III<sup>7</sup>, Brock A. Harpur<sup>8</sup>, Panagiotis  
4 Ioannidis<sup>9</sup>, Beryl M. Jones<sup>10</sup>, Clement F. Kent<sup>11</sup>, Livio Ruzzante<sup>12,13</sup>, Laura Sloofman<sup>7</sup>, Eckart  
5 Stolle<sup>14</sup>, Robert M. Waterhouse<sup>12,13</sup>, Amro Zayed<sup>11</sup>, Guojie Zhang<sup>3,4,5</sup>, William T. Wcislo<sup>2</sup>

6 <sup>1</sup>Department of Biology, Utah State University, Logan, UT 84322, U.S.A.

7 <sup>2</sup>Smithsonian Tropical Research Institute, Panama City, Republic of Panama

8 <sup>3</sup>State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology,  
9 Chinese Academy of Sciences, 650223, Kunming, China

10 <sup>4</sup>China National Genebank, BGI-Shenzhen, 518083, Shenzhen, Guangdong, China

11 <sup>5</sup>Centre for Social Evolution, Department of Biology, Universitetsparken 15, University of  
12 Copenhagen, DK-2100, Copenhagen, Denmark

13 <sup>6</sup>The Francis Crick Institute, London NW1 1AT, United Kingdom

14 <sup>7</sup>Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign  
15 (UIUC), Urbana, IL, USA.

16 <sup>8</sup>Donnelly Centre, University of Toronto, Toronto, Ontario, M3J 1P3, Canada

17 <sup>9</sup>Foundation for Research and Technology Hellas, Institute of Molecular Biology and  
18 Biotechnology, 70013 Vassilika Vouton, Heraklion, Greece

19 <sup>10</sup>Program in Ecology, Evolution, and Conservation Biology, University of Illinois at Urbana-  
20 Champaign, Urbana, IL 61801 U.S.A.

21 <sup>11</sup>Department of Biology, York University, Toronto, Ontario, M3J 1P3, Canada

22 <sup>12</sup>Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

23 <sup>13</sup>Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

24 <sup>14</sup>Institute of Biology, Martin-Luther-University Halle-Wittenberg, 06120 Halle, Germany

25

26

27 **Data Availability**

28 Genome assembly: PRJNA494873

29 Raw sequencing data for population genomics: PRJNA495036

30

31

32 **Running title**

33 Genome of the solitary alkali bee

34

35 **Key words**

36 Solitary bee, alternative pollinators, transposable elements, population genetics, sociogenomics

37

38 **Corresponding author**

39 Karen M. Kapheim

40 Department of Biology

41 Utah State University

42 5305 Old Main Hill

43 Logan, UT 84322-5305, USA

44 +1-435-797-0685

45 [karen.kapheim@usu.edu](mailto:karen.kapheim@usu.edu)

46

47

48

## 49 **ABSTRACT**

50 Alkali bees (*Nomia melanderi*) are solitary relatives of the halictine bees, which have  
51 become an important model for the evolution of social behavior, but for which few solitary  
52 comparisons exist. These ground-nesting bees defend their developing offspring against  
53 pathogens and predators, and thus exhibit some of the key traits that preceded insect sociality.  
54 Alkali bees are also efficient native pollinators of alfalfa seed, which is a crop of major economic  
55 value in the United States. We sequenced, assembled, and annotated a high-quality draft genome  
56 of 299.6 Mbp for this species. Repetitive content makes up more than one-third of this genome,  
57 and previously uncharacterized transposable elements are the most abundant type of repetitive  
58 DNA. We predicted 10,847 protein coding genes, and identify 479 of these undergoing positive  
59 directional selection with the use of population genetic analysis based on low-coverage whole  
60 genome sequencing of 19 individuals. We found evidence of recent population bottlenecks, but  
61 no significant evidence of population structure. We also identify 45 genes enriched for protein  
62 translation and folding, transcriptional regulation, and triglyceride metabolism evolving slower  
63 in alkali bees compared to other halictid bees. These resources will be useful for future studies of  
64 bee comparative genomics and pollinator health research.

65

## 66 **INTRODUCTION**

67 The comparative method is required for sociogenomics research, which aims to explain  
68 how social behavior evolves from a molecular perspective within the context of Darwinian  
69 evolution (Robinson *et al.* 2005). Eusociality is a special form of social behavior in animals that  
70 involves extreme levels of cooperation at the level of the group, manifest as queens and workers  
71 who distribute tasks related to reproduction, brood care, nest maintenance, and defense within a

72 colony (Wilson 1971). A large amount of comparative genomics research has focused on the  
73 insect order Hymenoptera, because ants, bees, and wasps display remarkable variation in social  
74 organization, and they represent at least five independent origins of eusociality in the past 200  
75 million years (Danforth *et al.* 2013; Branstetter *et al.* 2017). The comparative method is most  
76 powerful for understanding social evolution when it includes closely related species that are  
77 representative of the solitary ancestor from which eusociality arose (Rehan and Toth 2015).  
78 However, the rate at which genomic resources have become available for social Hymenoptera  
79 has far out-paced that for solitary species. Genome assemblies are publicly available for just  
80 three solitary bees and no solitary vespid wasps, compared to over 30 reference genomes  
81 currently available for social bees, wasps, and ants (Branstetter *et al.* 2018). This is in stark  
82 disproportion to the species that express solitary behavior among bees and wasps, most of which  
83 lead solitary lifestyles (Wcislo and Fewell 2017).

84 Alkali bees (*Nomia melanderi*) belong to the subfamily Nomiinae (Halictidae), a taxon  
85 composed of species that are solitary, though some express communal behavior and other forms  
86 of social tolerance (Wcislo and Engel 1996). The subfamily is the sister clade to the Halictinae,  
87 which includes both solitary and social lineages (Danforth *et al.* 2008). The alkali bees may be  
88 representative of the solitary ancestor from which eusociality likely evolved within the bee  
89 family Halictidae, and provide important phylogenetic context to comparative genomics (Brady  
90 *et al.* 2006; Gibbs *et al.* 2012). Alkali bees also possess several of the characteristic traits thought  
91 to be important in the ancestor of social halictids, including nest defense and other forms of  
92 maternal care (Batra and Bohart 1969; Batra 1970, 1972) (Fig. 1A). As such, this species has  
93 become an important model for testing hypotheses for the origins of eusociality, and has  
94 provided meaningful insight into the reproductive physiology of solitary bees (Kapheim 2017;

95 Kapheim and Johnson 2017a, 2017b). Development of genomic resources for this species will  
96 enable additional hypothesis testing regarding the solitary antecedents of eusociality in this  
97 family, and insects in general.

98         The development of genomic resources for alkali bees will also have practical and  
99 applied benefits. Alkali bees are native pollinators of alfalfa seed, which is a multi-billion dollar  
100 industry in the United States, accounting for one-third of the \$14 billion value attributed to U.S.  
101 bee-pollinated crops (Van Deynze *et al.* 2008; U.S. Department of Agriculture 2014). With  
102 issues of honey bee health and colony loss over the last decade, increased attention has been  
103 placed on the need to find alternative pollinators for many of our most important crops.  
104 Aggregations of alkali bees have been sustainably managed alongside alfalfa fields in  
105 southeastern Washington state for several decades (Cane 2008), and they are more effective  
106 pollinators of this crop than honey bees (Batra 1976; Cane 2002). Moreover, as a naturally  
107 aggregating native species, they are less costly pollinators than alfalfa leafcutter bees (*Megachile*  
108 *rotundata*), which must be purchased commercially (James 2011). Genomic resources have been  
109 an invaluable resource for the study of honey bee health and management, and are thus likely to  
110 benefit this important pollinator as well.

111         Here we present a draft genome assembly and annotation for *N. melanderi*, along with  
112 initial genomic comparisons with other Hymenoptera, a description of transcription factor binding  
113 sites, and population genetic analyses based on resequencing of individuals from throughout the  
114 southeastern Washington population. These resources will provide an important foundation for  
115 future research in sociogenomics and pollinator health.

116

## 117 **MATERIALS AND METHODS**

## 118 **Genome sequencing and assembly**

119 ***Sample collections:*** All of the bees used for sequencing were collected from nesting  
120 aggregations in and around Touchet, Washington (USA) with permission from private land  
121 owners in June 2014 or June 2015. Adult males and females were captured live, and flash frozen  
122 in liquid nitrogen. They were transported in a dry nitrogen shipper, and then stored at -80 °C  
123 until nucleic acid extraction.

124 ***DNA and RNA isolation:*** For genome sequencing, we isolated genomic DNA from  
125 individual males in three separate reactions targeting either the head or one half of a thorax. We  
126 used a Qiagen MagAttract kit, following the manufacturer's protocol, with two 200 µl elutions in  
127 AE buffer. We isolated RNA from three adult females using a Qiagen RNeasy kit, following the  
128 manufacturer's protocol, eluting once in 50 µl of water. We extracted RNA from the head and  
129 rest of the body separately for each female. For whole genome resequencing, we isolated  
130 genomic DNA of 18 adult females and one male from half of a thorax with a Qiagen MagAttract  
131 kit, as above. DNA was quantified with a dsDNA high sensitivity Qubit reaction, and quality was  
132 assessed on an agarose gel. RNA was quantified on a Nanodrop spectrophotometer, and quality  
133 was assessed with a Bioanalyzer.

134 ***Sequencing:*** All library preparation and sequencing was performed at the Roy J. Carver  
135 Biotechnology Center at University of Illinois at Urbana-Champaign. Two shotgun libraries  
136 (350-450 bp, 500-700 bp) were prepared from the DNA of a single haploid male with the Hyper  
137 Kapa Library Preparation kit (Kapa Biosystems). Three mate-pair libraries (3-5 kb, 8-10 kb, 15-  
138 20 kb) were constructed from DNA pooled from five individual males using the Nextera Mate  
139 Pair Library Sample Prep kit (Illumina, CA), followed by the TruSeq DNA Sample Prep kit. A

140 single RNA library was constructed from pooled RNA from the six female tissue samples with  
141 the TruSeq Stranded mRNA Library Construction kit (Illumina, CA).

142 DNA libraries were quantitated by qPCR and sequenced on a HiSeq2500 for 251 cycles  
143 from each end of the fragments using a TruSeq Rapid SBS kit version 2. Shotgun libraries were  
144 sequenced on a single lane, and mate-pair libraries were pooled and sequenced on a single lane.  
145 RNA libraries were sequenced on a single lane for 161 cycles from each end of the fragments.  
146 Fastq files were generated and demultiplexed with the bcl2fastq v1.8.4 Conversion Software  
147 (Illumina).

148 **Genome assembly:** The DNA shotgun and mate-pair library sequencing generated a total  
149 of 593,526,700 reads. After adapter trimming, these reads were filtered for quality (Phred  $64 < 7$ )  
150 and excessive ( $\geq 10$ ) Ns. We removed PCR duplicates from read pairs.

151 We used SOAPdenovo 2 with default parameters for genome assembly. We began by  
152 constructing contigs from the shotgun library reads split into kmers, which were used to  
153 construct a de Bruijn graph. Filtered reads were then realigned onto the contigs, and used to  
154 construct scaffolds based on shared paired-end relationships between contigs. We then closed  
155 gaps in the assembly using information from paired-end reads that mapped to a unique contig  
156 and a gap region.

157 **BUSCO assessment of assembly completeness:** The genome assembly completeness in  
158 terms of expected gene content was quantified using the Benchmarking Universal Single-Copy  
159 Ortholog (BUSCO) assessment tool (Waterhouse *et al.* 2018) for *N. melanderi* and seven other  
160 Apoidea species. Assembly completeness assessments employed BUSCOv3.0.3 with Augustus  
161 3.3 (Stanke *et al.* 2006), HMMER 3.1b2 (Finn *et al.* 2011), and BLAST+ 2.7.1 (Camacho *et al.*

162 2009) (Camacho et al. 2009), using both the hymenoptera\_odb9 and the insecta\_odb9 BUSCO  
163 lineage datasets and the Augustus species parameter ‘honeybee1’.

## 164 **Genome annotation**

165 **Gene annotation:** We predicted gene models based on homology and *de novo* methods.  
166 Results were integrated with GLEAN (Elsik *et al.* 2014). Homology based gene prediction used  
167 the gene models of four species (*Apis mellifera*, *Acromyrmex echinator*, *Drosophila*  
168 *melanogaster*, and *Homo sapiens*). We used TBLASTN to gather a non-redundant set of protein  
169 sequences, and then selected the most similar proteins for each candidate protein coding region  
170 based on sequence similarity. Short fragments were connected with a custom script (SOLAR),  
171 and Genewise (v2.0) (Birney *et al.* 2004) was used to generate the gene structures based on the  
172 homology alignments. This generated four gene sets, based on homology with four different  
173 species.

174 We used Augustus (Stanke *et al.* 2006) and SNAP (Johnson *et al.* 2008) for *de novo* gene  
175 prediction, with parameters trained on 500-1,000 intact genes from the homology-based  
176 predictions. We chose genes that were predicted by both programs for the final *de novo* gene set.

177 The four homology-based gene sets and one *de novo* gene set were integrated to generate  
178 a consensus gene set with GLEAN. We then filtered genes affiliated with repetitive DNA and  
179 genes whose CDS regions contained more than 30% Ns. Repetitive DNA was identified through  
180 annotation of tandem repeats (Tandem Repeats Finder v4.04) (Benson 1999) and transposable  
181 elements (TEs). This initial identification of TEs was performed based on homology-based and  
182 *de novo* predictions. For the homology-based approach, we used RepeatMasker (v3.2.9) and  
183 RepeatProteinMask (v3.2.9) (“Smit AFA, Hubley R, Green P: RepeatMasker. Available at:  
184 <http://www.repeatmasker.org>. [Accessed 9 April 2013]”) against a custom build of the Repbase



185 library. *De novo* predictions were performed with LTR\_FINDER (v1.0.5) (Xu and Wang 2007),  
186 PILER (v1.0) (Edgar and Myers 2005), and RepeatScout (v1.0.5) (Price *et al.* 2005). Results  
187 were used as an input library for a second run of RepeatMasker.

188 We used the 571,457,212 reads generated from RNA sequencing to polish the gene set.  
189 After filtering, we mapped reads to the genome with TopHat (Trapnell *et al.* 2009), and used  
190 Cufflinks (Trapnell *et al.* 2012) to assemble transcripts. Assembled transcripts were then used to  
191 predict ORFs. Transcript-based gene models with intact ORFs that had no overlap with the  
192 GLEAN gene set were added. GLEAN gene models were replaced by transcript-based gene  
193 models with intact ORFs when there was a discrepancy in length or merging of gene models.  
194 Transcripts without intact ORFs were used to extend the incomplete GLEAN gene models to  
195 find start and stop codons.

196 Putative gene functions were assigned to genes based on best alignments to the Swiss-  
197 Prot database (Release 2013\_11) (Bairoch 2004) using BLASTP. We used InterPro databases  
198 v32.0 (Zdobnov and Apweiler 2001; Quevillon *et al.* 2005) including Pfam, PRINTS, PROSITE,  
199 ProDom, and SMART to identify protein motifs and domains. Gene Ontology terms were  
200 obtained from the corresponding InterPro entries.

201 ***BUSCO assessment of annotation completeness:*** Annotated gene set completeness in  
202 terms of expected gene content was quantified using the BUSCO assessment tool (Waterhouse *et*  
203 *al.* 2018) for *N. melanderi* and seven other Apoidea species. Gene sets were first filtered to select  
204 the single longest protein sequence for any genes with annotated alternative transcripts. Gene set  
205 completeness assessments employed BUSCOv3.0.3 with HMMER 3.1b2 (Finn *et al.* 2011), and  
206 BLAST+ 2.7.1 (Camacho *et al.* 2009), using both the hymenoptera\_odb9 and the insecta\_odb9  
207 BUSCO lineage datasets.

208           **Transcription factor motif scans:** We generated binding scores for 223 representative  
209 transcription factor (TF) binding motifs in the *N. melanderi* genome. Motifs representative of TF  
210 clusters with at least one ortholog in bees (Kapheim *et al.* 2015) were selected from  
211 FlyFactorSurvey (Zhu *et al.* 2011). After masking tandem repeats with Tandem Repeat Finder,  
212 we produced normalized genome-wide scoring profiles for each selected TF motif in the genome  
213 based on sliding windows of 500 bp with 250 bp overlap. We used the HMM-based motif  
214 scoring program Stubb (Sinha *et al.* 2006) with a fixed transition probability of 0.0025 and a  
215 background state nucleotide distribution learned from 5 kb regions without coding features of  
216 length > 22 kb. We then normalized these motif scores using two different methods. First, we  
217 created a “Rank Normalized” matrix, to normalize the window scores across each motif on a  
218 scale of 0 (best) to 1 (worst). Second, we created a “G/C Normalized” matrix, by considering  
219 each window’s GC content. Motifs with high GC content are likely to produce a high Stubb  
220 score in a GC rich window. We thus separated genomic windows into 20 bins of equal size based  
221 on GC content, and performed rank-normalization separately within each bin. We next  
222 summarized motif scores at the gene level. For each gene, we calculated a score for each motif as  
223  $P_{gm} = 1 - (1 - Ngm)^{Wg}$ , where  $Ngm$  is the best normalized score for motif  $m$  among the  $Wg$   
224 windows that fall within the regulatory region of the gene  $g$ . We defined the regulatory region of  
225 the gene in five different ways: *5Kup2Kdown* – 5000 bp upstream to 2000 bp downstream of a  
226 gene’s transcription start site (TSS), *5Kup* – 5000 bp upstream of a gene’s TSS, *1Kup* – 1000 bp  
227 upstream of a gene’s TSS, *NearStartSite* – all genomic windows that are closer to the gene’s TSS  
228 than any other gene TSS, *GeneTerr* – all genomic windows between the boundary positions of  
229 the nearest non-overlapping gene neighbors within at least 5000 bp upstream of the TSS.

230 We used the results of these target motif scans to check for transcription factor motif  
231 enrichment among gene sets of interest (i.e., genes under selection). For each normalization  
232 method and regulatory region, we created two motif target gene sets: a "conservative" set that  
233 contains only the top 100 genes by normalized score and a more "liberal" set that contains the  
234 800 top genes. Enrichment tests for genes of interest were performed using the one-sided Fisher  
235 exact test for each of 1784 motif target sets defined using the two thresholds, both "G/C" and  
236 "Rank" normalization procedures, the *IKup* (likely the core promoter) and *GeneTerr* (likely  
237 containing distal enhancers) regulatory region definitions, and each of the representative 223  
238 motifs. Multiple hypothesis test corrections were performed using the Benjamini-Hochberg  
239 procedure (Benjamini and Hochberg 1995). For significantly enriched motifs (adjusted- $p < 6E-$   
240  $04$ ), we determined if an ortholog of the fly transcription factor protein was present in the *N.*  
241 *melanderi* genome using blastp with e-value  $< 10e-3$  and % identity  $\geq 50$ .

242 ***Transposable element identification:*** We performed a more detailed *de novo*  
243 investigation of transposable elements in the *N. melanderi* genome using raw sequencing reads in  
244 a genome assembly-independent approach. First, we filtered a subset of five million raw reads  
245 for mitochondrial contamination to avoid biasing the detection of highly repetitive sequences.  
246 This involved aligning reads to the genome assembly with bwa-mem (Li 2013), and evaluating  
247 read depth with bedtools (Quinlan and Hall 2010). We identified contigs and scaffolds with high  
248 coverage ( $\geq 500x$ ) as potential mitochondrial sequences, based on the assumption that the  
249 number of sequenced mitochondrial copies is much higher than that of the nuclear genome.  
250 These contigs and scaffolds were further analyzed for sequence similarity (blastn v. 2.2.28+) to  
251 the mitogenome of the closest available bee species, *Halictus rubicundus* (KT164656.1). We  
252 identified five scaffolds as putatively mitochondrial (scaffold235256, scaffold241193,

253 scaffold252191, scaffold252994, scaffold257806). Reads aligning to these scaffolds were filtered  
254 from the analysis.

255 The remaining reads were used for repeat analysis in five iterations of the transposable  
256 element discovery program DnaPipeTE v1.1 (Goubert *et al.* 2015), following Stolle *et al.* (2018).  
257 Each iteration used a new set of the same number of reads randomly sampled from the filtered  
258 reads. The analysis was repeated for different number of reads to represent a genome sequence  
259 assembly length coverage of 0.20x-0.40x in steps of 0.05x. This series of repeat content  
260 estimates determines the amount of input data that provides a stable estimate of genomic repeat  
261 content, and thus ensures that adequate coverage has been obtained for accurate estimates. The  
262 final set of repetitive elements was generated based on 0.30x coverage, using RepeatMasker  
263 v4.0.7 and a 10% sequence divergence cut-off. Overlap between repetitive element annotations  
264 and genes was detected with bedtools.

### 265 **Orthology delineation**

266 Orthologous groups (OGs) delineated across 116 insect species were retrieved from  
267 OrthoDB v9.1 (Zdobnov *et al.* 2017) to identify orthologs. The OrthoDB orthology delineation  
268 procedure employs all-against-all protein sequence alignments to identify all best reciprocal hits  
269 (BRHs) between genes from each pair of species. It then uses a graph-based approach that starts  
270 with BRH triangulation to build OGs containing all genes descended from a single gene in the  
271 last common ancestor of the considered species. The annotated proteins from the genomes of *N.*  
272 *melanderi* were first filtered to select one protein-coding transcript per gene and then mapped to  
273 OrthoDB v9.1 at the Insecta level, using all 116 species and an unpublished halictid bee genome  
274 (*Megalopta genalis*; Kapheim *et al.* unpublished) for orthology mapping. The OrthoDB

275 orthology mapping approach uses the same BRH-based procedure as for building OGs, but only  
276 allowing proteins from the mapped species to join existing OGs.

## 277 **Phylogenomic analysis**

278 We reconstructed a molecular species phylogeny from 2,025 universal single-copy  
279 orthologs among the protein sequences of 15 insects including *N. melanderi* (Table S1-S2). The  
280 protein sequences from each orthogroup were first aligned with Muscle 3.8.31 (Edgar 2004),  
281 then trimmed to retain only confidently aligned regions with TrimAl v1.3 (Capella-Gutierrez *et*  
282 *al.* 2009), and then concatenated to form the 15 species superalignment of 688,354 columns. The  
283 maximum likelihood phylogeny was then estimated using RAxML 8.0.0 (Stamatakis 2014), with  
284 the PROTGAMMAJTT substitution model, setting the body louse (*Pediculus humanus*) as the  
285 outgroup species, and performing 100 bootstrap samples to obtain support values.

286 With these data, we performed a comparative orthology analysis to identify genes with  
287 universal, widely shared, or lineage-specific/restricted distributions across the selected species,  
288 or with identifiable orthologs from other insect species from OrthoDB v9.1. Ortholog presence,  
289 absence, and copy numbers were assessed for all OGs across the 15 species to classify genes  
290 according to their orthology profiles. The categories (each mutually exclusive) included: 1)  
291 Single-copy in all 15 insect species; 2) Present in all 15 insect species; 3) Halictidae: Present in  
292  $\geq 2$  Halictidae but none of the other 11 species; 4) Apidae + Mrot: Present in  $\geq 2$  Apidae and  
293 *Megachile rotundata* but none of the other 11 species; 5) Apoidea: Present in  $\geq 1$  Halictidae,  
294 present in  $\geq 1$  Apidae and *Megachile rotundata* but none of the other 7 species; 6) Formicoidea:  
295 Present in  $\geq 2$  Formicoidea but none of the other 11 species; 7) Apoidea + Formicoidea: Present  
296 in  $\geq 2$  Apoidea, present in  $\geq 1$  Formicoidea but not in *Polistes dominula* or *Cephus cinctus* or  
297 *P. humanus*; 8) Pdom + Ccin: Present in *P. dominula* and *C. cinctus* but none of the other 13

298 species; 9) Pdom or Ccin or Phum: Present in  $\geq 2$  of *P. dominula* or *C. cinctus* or *P. humanus*  
299 and none of the other 12 species; 10) Hymenoptera: Present in  $\geq 2$  Hymenoptera and absent  
300 from *P. humanus*; 11) Present < 13: Present in <13 of the 15 species, i.e., a patchy distribution  
301 not represented by any other category; 12) Other orthology: Present in any other insect from  
302 OrthoDB v9.1; 13) No orthology: No identifiable orthology at the OrthoDB v9.1 Insecta level.

### 303 **Population genetic analysis**

304 ***SNP discovery and filtering:*** We used sequences generated from the 18 females and one  
305 male to characterize genetic variants following GATK best practices  
306 (<https://software.broadinstitute.org/gatk/best-practices/>). Reads were pre-processed by quality  
307 trimming using sickle with default parameters (Joshi and Fass 2011). We then converted paired  
308 reads to BAM format and marked adapters with Picard tools (“Picard.  
309 <http://picard.sourceforge.net/>. Accessed 12 January 2016”). Reads were aligned to the genome  
310 with bwa-mem wrapped through Picard tools (CLIPPING\_ATTRIBUTE=XT,  
311 CLIPPING\_ACTION=2, INTERLEAVE=true, NON\_PF=true). Alignments were then merged  
312 with MergeBamAlignment (CLIP\_ADAPTERS=false, CLIP\_OVERLAPPING\_READS=true,  
313 INCLUDE\_SECONDARY\_ALIGNMENTS=true, MAX\_INSERTIONS\_OR\_DELETIONS=-1,  
314 PRIMARY\_ALIGNMENT\_STRATEGY=MostDistant, ATTRIBUTES\_TO\_RETAIN=XS).  
315 PCR duplicates were marked with the function MarkDuplicatesWithMateCigar  
316 (OPTICAL\_DUPLICATE\_PIXEL\_DISTANCE=2500, MINIMUM\_DISTANCE=300). We next  
317 identified and realigned around indels using the Picard tools functions RealignerTargetCreator  
318 and IndelRealigner.

319 We performed variant calling in two rounds. The first pass was to generate a high quality  
320 SNP set that could be used for base quality recalibration, followed by a second pass of variant

321 calling. For both rounds, we used the HaplotypeCaller function in Picard tools (--  
322 variant\_index\_type LINEAR, --variant\_index\_parameter 128000, -ERC GVCF), followed by  
323 joint genotyping for the 18 females and individual genotyping for the male sample  
324 (GenotypeGVCFs). Haplotype caller was run set with ploidy level = 2n for all samples,  
325 including the haploid male. The latter was used to identify low-confidence or spurious SNPs that  
326 could be filtered from the female calls.

327 Variant filtering followed the GATK generic recommendations (--filterExpression "QD <  
328 2.0, FS > 60.0, MQ < 40.0, ReadPosRankSum < -8.0, --restrictAllelesTo BIALLELIC). These  
329 were further filtered for SNPs identified as heterozygous in the male sample and for which  
330 genotypes were missing in any sample (--max-missing-count 0).

331 This set of high-confidence SNPs was used as input for base quality score recalibration  
332 for the 18 females. The second round of variant calling and filtering for these samples followed  
333 that of the first round, with the exception that we allowed missing genotypes in up to 8 samples.  
334 We then applied a final, more stringent set of filters using vcftools (Danecek *et al.* 2011) (--min-  
335 meanDP 5, --max-missing-count 4, --maf 0.05, --minGQ 9, --minDP 3). This yielded a final set  
336 of 412,800 high confidence SNPs used in the downstream analyses (File S1).

337 **Structure analysis:** We evaluated the potential for population structure by estimating  
338 heterozygosity, relatedness, and Hardy-Weinberg disequilibrium within our samples using  
339 vcftools. We also used ADMIXTURE v.1.3 (Alexander *et al.* 2009) to look for evidence of  
340 population structure (N=18 diploids). We randomly extracted SNPS that were at least 1000bp  
341 apart across the genome and ran K = 1-4 for three independent datasets.

342           **SNP function:** We identified the functional role (e.g., upstream, synonymous, non-  
343 synonymous, etc.) of SNPs using SNPEFF (Cingolani *et al.* 2012) for all SNPs within our data  
344 set (N = 412,800).

345           **Genetic diversity:** We characterized genetic diversity by evaluating pi and Tajima's D in  
346 10Kb and 1Kb windows with vcftools (--window-pi, --TajimaD, --site-pi). We mapped gene  
347 models to these windows with bedtools intersect, and Tajima's D and pi values were averaged  
348 over each gene model using the aggregate function in R (Team 2016). We then calculated the  
349 cumulative percentile for pi and Tajima's D for each gene using the ecdf function in R. These  
350 percentiles were then multiplied and recalculated. Genes for the joint percentile of pi and  
351 Tajima's D that fell in the lowest 5% were considered to be under ongoing positive selection. To  
352 estimate genetic diversity across the genome in windows, we first calculated coverage at each  
353 site within 1Kb windows across the genome using bedtools coverage. Within each window, we  
354 estimated the proportion of sites with at least 5 reads of coverage. We used this value as the  
355 denominator to calculate pi within 1Kb windows.

356           **Effective population size and demography:**

357           We estimated Ne using SMC++ (Terhorst *et al.* 2017). We randomly selected 4 large  
358 scaffolds (> 1 Kb) and estimated effective population size of our single *Nomia melanderi*  
359 population from 1000 to 100000 years before present. We assumed a single generation per year  
360 and a mutation rate of  $6.8 \times 10^{-9}$  (Liu *et al.* 2017). For each scaffold, we created 6 datasets by  
361 randomly selecting between 5 and 8 individuals without replacement. We used these files to  
362 estimate Ne using the cross-validation for each scaffold.

363           We evaluated the possibility of recent demographic changes by estimating Tajima's D in  
364 1000bp windows across the genome for all samples (Tajima 1989).



365 **Evolutionary rate analysis**

366 Single copy orthologs were extracted from OGs identified above for *Lasioglossum*  
367 *albipes*, *Dufourea novaengliae*, *M. genalis*, and *N. melanderi*. Peptide alignments were obtained  
368 by running GUIDANCE2 (Penn *et al.* 2010) with the PRANK aligner (Löytynoja 2014) and  
369 species tree ((Dnov:67.51,(Nmel:58.18,(Mgen:47.03,Lalb:47.03):11.15):9.33); (Branstetter *et al.*  
370 2017)) on each orthogroup. Low scoring residues (scores < 0.5) were masked to N using  
371 GUIDANCE2 to mask poor quality regions of each alignment. PAL2NAL (Suyama *et al.* 2006)  
372 was used to back-translate aligned peptide sequences to CDS and format alignments for PAML.  
373 PAML (Yang 2007) was run to evaluate the likelihood of multiple hypothesized branch models  
374 of dN/dS relative to two null models with trees and parameters as follows:  
375 M0: (Dnov:67.51,(Nmel:58.18,(Mgen:47.03,Lalb:47.03):11.15):9.33); (model = 0, fix\_omega =  
376 0, omega = 0.2; all branches same omega)  
377 M1a: (Dnov:67.51,(Nmel:58.18 #1 ,(Mgen:47.03,Lalb:47.03):11.15):9.33);  
378 (model = 2, fix\_omega = 1, omega = 1; neutral evolution for Nmel branch)  
379 M2a: (Dnov:67.51,(Nmel:58.18 #1,(Mgen:47.03,Lalb:47.03):11.15):9.33); (model = 2,  
380 fix\_omega = 0, omega = 0.2; Nmel branch different omega)  
381 M1b: (Dnov:67.51,(Nmel:58.18,(Mgen:47.03 #1,Lalb:47.03):11.15):9.33); (model=2,  
382 fix\_omega=1, omega = 1; neutral evolution for Mgen branch)  
383 M2b: (Dnov:67.51,(Nmel:58.18,(Mgen:47.03 #1,Lalb:47.03):11.15):9.33); (model=2,  
384 fix\_omega=0, omega=0.2; Mgen branch different omega)  
385 M1c: (Dnov:67.51,(Nmel:58.18,(Mgen:47.03,Lalb:47.03 #1):11.15):9.33); (model=2,  
386 fix\_omega=1, omega=1; neutral evolution for Lalb branch)

387 M2c: (Dnov:67.51,(Nmel:58.18,(Mgen:47.03,Lalb:47.03 #1):11.15):9.33); (model=2,  
388 fix\_omega=0, omega=0.2; Lalb branch different omega)  
389 Orthogroups with  $dS > 2$  were removed, and likelihood ratio tests were performed to determine  
390 the most likely value of omega for each branch.

### 391 **Functional Enrichment Tests**

392 We performed all tests of functional enrichment using the GOstats package (Gentleman  
393 and Falcon 2013) in R version 3.4.4. We used terms that were significantly enriched ( $p < 0.05$ ) to  
394 build word clouds with the R packages tm (Feinerer *et al.* 2008), SnowballC (Bouchet-Valat  
395 2014), and wordcloud (Fellows 2018).

### 396 **Data Availability**

397 Sequence data are available at NCBI (BioProject PRJNA495036). The genome assembly  
398 is available at NCBI (BioProject PRJNA494873). Genetic variants and genotypes are available  
399 in VCF format in File S1. TF binding motif scores are in File S2. Repetitive DNA content is in  
400 File S3. SNP effects are in File S4. The genome annotation (GFF format) is in File S5. All  
401 supplementary tables (Table S1-S8) and files (Files S1-S5) have been deposited at FigShare.

402

## 403 **RESULTS AND DISCUSSION**

404 The *N. melanderi* genome assembly resulted in 268,376 scaffolds (3,194 > 1 kb) with an  
405 N50 scaffold length of 2.05 Mb (Table 1). Total size is estimated to be 299.6 Mb, based on a k-  
406 mer analysis with  $k = 17$  and a peak depth of 70. CEGMA analysis indicated 244 of 248  
407 (98.39%) core eukaryotic genes were completely assembled, and 10.25% of the detected  
408 CEGMAs had more than one ortholog. BUSCO analyses indicated 98.8% of Insecta BUSCOs  
409 were complete in the assembly (Table S3).

410

411 **Table 1** Comparison of genome assemblies among bees, including *Nomia melanderi*.

Species	Genome size (Mb)	Number scaffolds	N50 Scaffold length	Predicted Genes	Coverage (X)	Reference
<i>Nomia melanderi</i>	299.6	268,376 (3,194 > 1kb)	2,054,768	10,847	75	---
<i>Lasioglossum albipes</i>	416	41,377	616,426	13,448	96	(Kocher <i>et al.</i> 2013)
<i>Dufourea novaeangliae</i>	291	84,187	2,397,596	12,453	133	(Kapheim <i>et al.</i> 2015)
<i>Megachile rotundata</i>	273	6,266	1,699,680	12,770	272	(Kapheim <i>et al.</i> 2015)
<i>Bombus impatiens</i>	248	5,559	1,399,493	15,896	108	(Sadd <i>et al.</i> 2015)

412

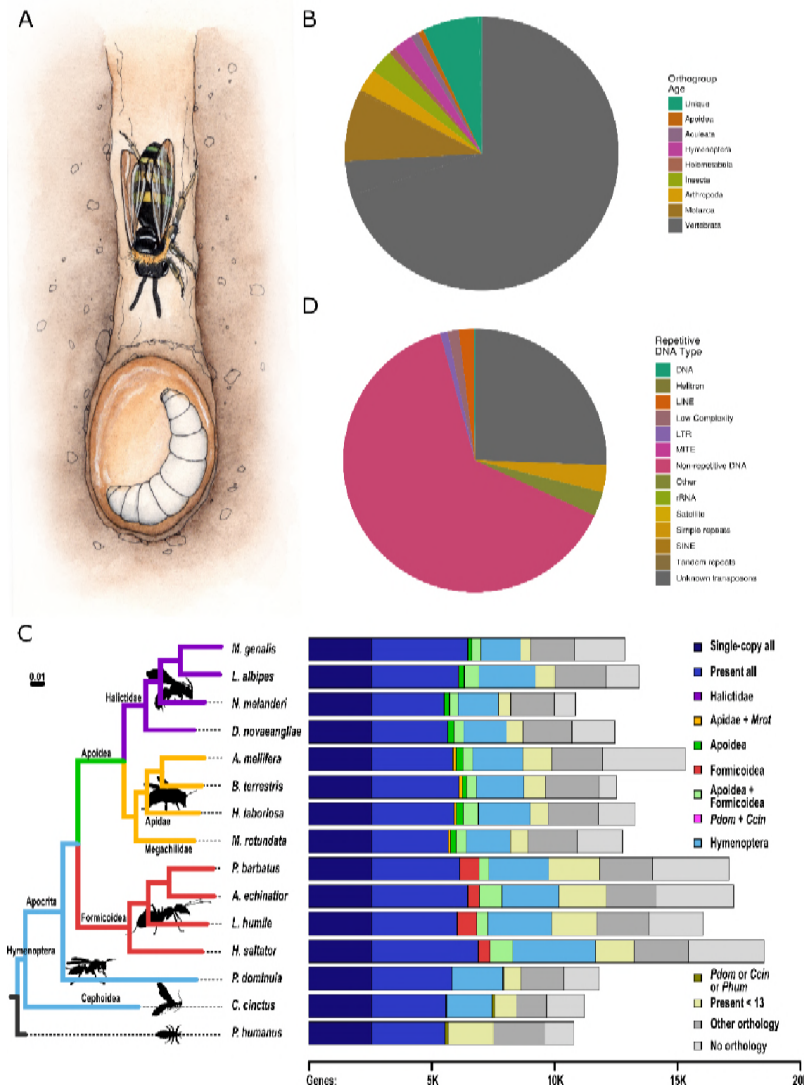
413 Our official gene set includes 10,847 predicted protein-coding gene models. This is likely  
414 to be a relatively complete gene set, as 96.0% of Insecta BUSCOs were identified as complete,  
415 which is comparable to other bee genomes (Table S3). Most (8,075) of the predicted genes  
416 belong to ancient OGs that include orthologs in vertebrate species. However, there were 819  
417 genes without any known orthologs (Fig. 1B). Our comparative analysis with representative  
418 Hymenoptera species and the outgroup, *P. humanus*, identified 2,025 single-copy orthologs from  
419 which we constructed the molecular species phylogeny that confidently places Halictidae as a  
420 sister group to the combined Apidae and Megachilidae groups within Apoidea (Fig. 1C).  
421 Orthology delineation showed that 92.2% of *N. melanderi* predicted genes have orthologs in  
422 other insects and only 16 of them were unique to the family Halictidae (Fig. 1C). Transcription  
423 factor motif binding scores for each gene are available in File S2.

424 In a genome-assembly independent approach using short reads and DnaPipeTE, we  
425 assembled 54,236 repetitive elements, suggesting that 37.5% of the *N. melanderi* genome is  
426 repetitive content (File S3; Fig. 1D). We identified transposable elements from all major groups  
427 (LTR, LINE, SINE, DNA, Helitron) and other elements with similarities to unclassified repeats

428 (7,866 total annotated repeats), but unknown elements are the most abundant type of transposon  
429 (25.5%) (Fig. 1D), showing no similarities to known repetitive elements, conserved domains, or  
430 sequences in NCBI's non-redundant nt database.

431       Of annotated transposable elements, LINE retrotransposons (most common: I and  
432 Jockey) were the most abundant, followed by LTR retrotransposons (most common: Gypsy) and  
433 small amounts of DNA (mostly Tc1-Mariner, PiggyBac, hAT and Kolobok) or other transposons  
434 (File S3). Some annotations suggest the presence of Crypton, Helitron and Maverick elements as  
435 well as 5S/tRNA SINE (File S3). A majority of the detected retroelements show little sequence  
436 divergence, indicating recent activity, particularly Gypsy (LTR), Copia (LTR), I (LINE) and R2  
437 (LINE).

438       Annotation of the genome assembly yielded 25.93 Mbp of masked sequences (8.59% at  
439 10% sequence divergence), which is less than the repetitive fraction of >37% inferred by  
440 DnaPipeTE. Even at a 20% sequence divergence threshold, only 43.36 Mbp (14.37%) were  
441 masked, suggesting that a substantial fraction of the repetitive part of the genome is not part of  
442 the genome assembly, likely due to the technical limitations in assembling repetitive elements  
443 from short reads.



444

445 **Figure 1** *Nomia melanderi* genome characteristics and comparative context. (A) *N. melanderi* are  
 446 ground-nesting bees with maternal care. (B) Most of the protein-coding genes belong to OGs that  
 447 include vertebrates or other metazoans, and are thus widely conserved. (C) *N. melanderi* species  
 448 phylogeny (left) and gene orthology (right). The maximum likelihood 15-species molecular  
 449 phylogeny estimated from the superalignment of 2,025 single-copy orthologs recovers supported  
 450 families. Branch lengths represent substitutions per site, all nodes achieved 100% bootstrap  
 451 support. Right: Total gene counts per species partitioned into categories from single-copy  
 452 orthologs in all 15 species, or present but not necessarily single-copy in all (i.e., including gene  
 453 duplications), to lineage-restricted orthologs (Halictidae, Apidae and *M. rotundata*, Apoidea,  
 454 Formicoidea, Apoidea and Formicoidea, Hymenoptera, specific outgroups), genes showing  
 455 orthology in less than 13 species (i.e., patchy distributions), genes present in the outgroups (present  
 456 in *P. domunila* or *C. cinctus*, present in *P. domunila* or *C. Cinctus* or *P. humanus*), and genes with  
 457 orthologs from other sequenced insect genomes or with no identifiable orthology. The purple  
 458 Halictidae bar is present but barely visible as only 16 to 32 orthologous genes were assigned to the  
 459 Halictidae-restricted category. (D) A large proportion of repetitive DNA consists of  
 460 uncharacterized transposable elements, but all major transposon groups were detected.

461

462 Our population genetic analysis indicated our population is panmictic. We did not find  
463 any evidence of population structure among our samples. Across all three datasets run through  
464 STRUCTURE, the lowest CV error was found for  $K = 1$  ( $CV = 0.68$ ) (Fig. 2A). Likewise,  
465 pairwise relatedness estimates based on the unadjusted  $A_{jk}$  statistic were close to 0 ( $-0.084 - -$   
466  $0.047$ ) for all females in our population (Yang *et al.* 2010).

467 Solitary bees are expected to have high genetic diversity and large effective population  
468 sizes (Romiguier *et al.* 2014), and recent census data suggests there are 17 million females  
469 nesting in our study population in the Touchet Valley (Washington, USA) (Cane 2008).  
470 However, we find several lines of evidence to suggest that effective population size of our *N.*  
471 *melanderi* population has declined in the recent past. First, our estimates of genetic diversity  
472 were surprisingly low. Three of the 18 females in our dataset had significantly higher  
473 homozygosity than expected ( $p < 0.05$ ). Genetic diversity ( $\pi$ ) across the genome in 1Kb  
474 windows (corrected for coverage, see Methods) was estimated to be 0.00153. This is  
475 intermediate to diversity previously estimated for *Apis mellifera* (0.0131, (Harpur *et al.* 2014))  
476 and *Bombus impatiens* (0.002, (Harpur *et al.* 2017)).

477 Second, the genome-wide average Tajima's D was significantly greater than 0 (one-way  
478 T-test; mean = 0.77 +/- 0.002 SE;  $p < 0.00001$ ) indicating a recent population decline.

479 Third,  $N_e$  is predicted to have declined within the last 10,000 years (Fig. 2B). In the last  
480 2,000 years,  $N_e$  has had a median of 12,554 individuals (range: 3,119-3,978,942). The long, slow  
481 population decline reflected in our samples corresponds to a period during which much of  
482 Washington state was underwater due to glacial flooding, known as the Missoula Floods. Our

483 study area, Touchet Valley, was under Lake Lewis during this time, and was thus uninhabitable  
484 for ground-nesting bees.

485 More recent fluctuations in  $N_e$  may reflect less catastrophic events. Seed growers have  
486 maintained large nesting areas (“bee beds”) for alkali bees within a 240 km<sup>2</sup> watershed that  
487 encompasses our sampling area for several decades (Cane 2008). Some of these bee beds are  
488 among the largest nesting aggregations ever recorded, at up to 278 nests per m<sup>2</sup>. However,  
489 survey data suggests there are large fluctuations in population size, as the population increased 9-  
490 fold over an eight year period (1999-2006) (Cane 2008). Records from individual bee beds  
491 reflect these fluctuations. For example, a bed that was started in 1973 grew from 550 nesting  
492 females to 5.3 million nesting females in 33 years (Johansen *et al.* 1978; Cane 2008). However,  
493 other beds were destroyed or abandoned for decades at a time, only to be recolonized later. A  
494 large population crash occurred in the 1990s, likely due to use of a new pesticide (Cane 2008),  
495 and flooding events have caused massive valley-wide reproductive failures (Stephen 2003). Our  
496 wide range of  $N_e$  estimates and signatures of genetic bottlenecks likely reflect these population  
497 fluctuations.

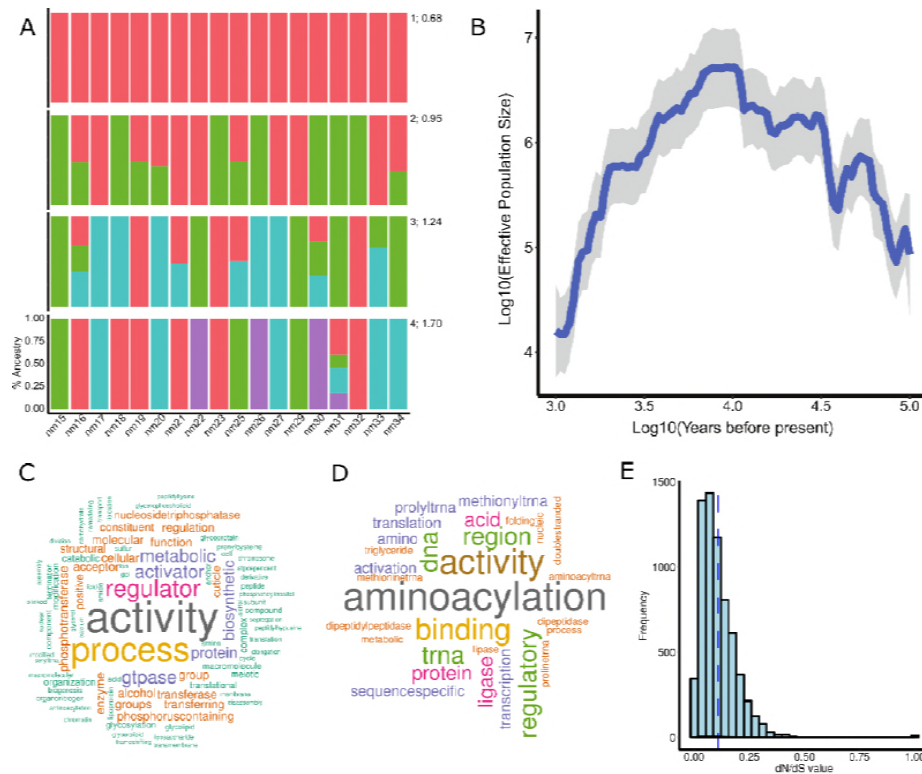
498 Our selection scan revealed 479 *N. melanderi* genes under positive directional selection.  
499 Genes under selection were highly conserved, and the age distribution was similar to the  
500 distribution across all predicted genes ( $\chi^2 = 54$ , d.f. = 48,  $p = 0.26$ ; Table S4). Genes showing  
501 signatures of ongoing positive selection were enriched for functions related to tRNA transfer and  
502 DNA/nucleosome binding (Fig. 2C, Table S5). Because DNA binding is typically an indicator of  
503 transcription factor activity, we performed enrichment analysis of genes under selection with our  
504 previously defined transcription factor motif target sets (File S2). The most enriched motif target  
505 sets (adjusted- $p < 6E-04$ ) included transcription factors involved in neural differentiation (*brick-*

506 *a-brack 1, prospero, nubbin, zelda, twin-of-eyeless, pox meso, worniu*) and neural secretory  
507 functions (*dimmed*) (Table S6). We identified 505,203 functional predictions for 412,800  
508 variable sites (SNPs) within 9,692 genes, most of which are intergenic (File S4).

509 Our analysis of evolutionary rates included 6,644 single-copy orthologs, most of which  
510 (95%) were evolving at similar rates across all four halictid bee lineages. We identified 61 *N.*  
511 *melanderi* genes that are evolving at a significantly different rate from other halictid bees (Table  
512 S7). Of these, the majority (74%) are evolving slower than in other lineages. These genes are  
513 significantly enriched for functions related to transcription and translation (Fig. 2D, Table S8).  
514 The distribution of estimated dN/dS values for *N. melanderi* genes was skewed toward zero, with  
515 a notable absence of values greater than one (Fig. 2E). This suggests that most genes in our  
516 analysis show evidence of neutral or purifying selection. This result is likely influenced by the  
517 vast evolutionary distance separating the four halictid lineages, which shared a common ancestor  
518 > 150 million years ago (Branstetter *et al.* 2017). Our set of single-copy orthologs was thus  
519 limited to highly conserved genes.

520 In conclusion, we present a high quality draft genome assembly of the solitary alkali bee,  
521 *N. melanderi*, that will be a valuable resource for both basic and applied research communities.





522

523 **Figure 2** *N. melanderi* population genetics. (A) Samples most likely originate from a single source  
 524 population. We tested for population structure for K=1-4 (right numbers) and found that the most  
 525 likely K = 1 (average CV error = 0.68 across three independent runs). K:CV is given to the right  
 526 of each row. (B) Estimates of  $N_e$  show evidence for a decline in effective population size in our  
 527 alkali bee population, beginning about 10,000 years before present. Blue line, median estimated  
 528  $N_e$ ; shaded gray area, 95% confidence intervals. (C) Genes under positive selection are  
 529 significantly enriched for molecular functions and biological processes related to tRNA transfer  
 530 and binding. (D) Genes with a slower evolutionary rate (dN/dS) in *N. melanderi* than in other  
 531 halictid bees are significantly enriched for processes and functions related to transcription and  
 532 translation. In B and C, the size of the word corresponds to the frequency to which that term  
 533 appears on a list of significantly enriched GO terms. (E) The distribution of dN/dS values for *N.*  
 534 *melanderi* genes are skewed toward zero, and none are greater than 1. Blue dashed line, mean  
 535 dN/dS.

536

## 537 REFERENCES

538 Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in  
 539 unrelated individuals. *Genome Res.* 19: 1655–64.

540 Bairoch, A., 2004 Swiss-Prot: Juggling between evolution and stability. *Brief. Bioinform.* 5: 39–

541 55.

- 542 Batra, S. W. T., 1970 Behavior of alkali bee, *Nomia-melanderi*, within nest (Hymenoptera-  
543 Halictidae). Ann. Entomol. Soc. Am. 63: 400–406.
- 544 Batra, S. W. T., 1976 Comparative efficiency of alfalfa pollination by *Nomia melanderi*,  
545 *Megachile rotundata*, *Anthidium florentinum* and *Pithitis smaragdula* (Hymenoptera:  
546 Apoidea). Source J. Kansas Entomol. Soc. 49: 18–22.
- 547 Batra, S. W. T., 1972 Some properties of the nest-building secretions of *Nomia*, *Anthophora*,  
548 *Hylaeus* and other bees. J. Kansas Entomol. Soc. 45: 208–218.
- 549 Batra, S. W., and G. E. Bohart, 1969 Alkali bees: response of adults to pathogenic fungi in brood  
550 cells. Science 165: 607.
- 551 Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and  
552 powerful approach to multiple testing. J R Stat Soc Ser B 57:.
- 553 Benson, G., 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids  
554 Res 27: 573–580.
- 555 Birney, E., M. Clamp, and R. Durbin, 2004 GeneWise and GenomeWise. Genome Res 14: 988–  
556 995.
- 557 Bouchet-Valat, M., 2014 Snowball stemmers based on the C libstemmer UTF-8 library [R  
558 package SnowballC version 0.5.1].
- 559 Brady, S. G., S. Sipes, A. Pearson, and B. N. Danforth, 2006 Recent and simultaneous origins of  
560 eusociality in halictid bees. Proc. R. Soc. B Biol. Sci. 273: 1643–1649.
- 561 Branstetter, M. G., A. K. Childers, D. Cox-Foster, K. R. Hopper, K. M. Kapheim *et al.*, 2018  
562 Genomes of the Hymenoptera. Curr. Opin. Insect Sci. 25: 65–75.
- 563 Branstetter, M. G., B. N. Danforth, J. P. Pitts, B. C. Faircloth, P. S. Ward *et al.*, 2017  
564 Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees.

- 565 Curr. Biol. 27: 1019–1025.
- 566 Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+:  
567 architecture and applications. BMC Bioinformatics 10: 421.
- 568 Cane, J. H., 2008 A native ground-nesting bee (*Nomia melanderi*) sustainably managed to  
569 pollinate alfalfa across an intensively agricultural landscape. Apidologie 39: 315–323.
- 570 Cane, J. H., 2002 Pollinating bees (Hymenoptera: Apiformes) of U.S. alfalfa compared for rates  
571 of pod and seed set. J. Econ. Entomol. 95: 22–27.
- 572 Capella-Gutierrez, S., J. M. Silla-Martinez, and T. Gabaldon, 2009 trimAl: a tool for automated  
573 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.
- 574 Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating  
575 and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome  
576 of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 6: 80–92.
- 577 Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format  
578 and VCFtools. Bioinformatics 27: 2156–2158.
- 579 Danforth, B. N., S. Cardinal, C. Praz, E. A. B. Almeida, and D. Michez, 2013 The impact of  
580 molecular data on our understanding of bee phylogeny and evolution. Annu. Rev. Entomol.  
581 58: 57–78.
- 582 Danforth, B. N., C. Eardley, L. Packer, K. Walker, A. Pauly *et al.*, 2008 Phylogeny of Halictidae  
583 with an emphasis on endemic African Halictinae. Apidologie 39: 86–101.
- 584 Van Deynze, A. E., S. Fitzpatrick, B. Hammon, M. H. McCaslin, D. H. Putnam *et al.*, 2008 Gene  
585 flow in alfalfa: biology, mitigation, and potential impact on production.:
- 586 Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high  
587 throughput. Nucleic Acids Res 32: 1792–1797.

- 588 Edgar, R. C., and E. W. Myers, 2005 PILER: identification and classification of genomic repeats.  
589 *Bioinformatics* 21: i152–i158.
- 590 Elsik, C. G., K. C. Worley, A. K. Bennett, M. Beye, F. Camara *et al.*, 2014 Finding the missing  
591 honey bee genes: Lessons learned from a genome upgrade. *BMC Genomics* 15: 1–29.
- 592 Feinerer, I., K. Hornik, and D. Meyer, 2008 Text mining infrastructure in R. *J. Stat. Softw.* 25:  
593 1–54.
- 594 Fellows, I., 2018 wordcloud: Word Clouds. R package version 2.6.
- 595 Finn, R. D., J. Clements, and S. R. Eddy, 2011 HMMER web server: interactive sequence  
596 similarity searching. *Nucleic Acids Res* 39: W29-37.
- 597 Gentleman, R., and S. Falcon, 2013 *Package “GOstats.”*
- 598 Gibbs, J., S. G. Brady, K. Kanda, and B. N. Danforth, 2012 Phylogeny of halictine bees supports  
599 a shared origin of eusociality for *Halictus* and *Lasioglossum* (Apoidea: Anthophila:  
600 Halictidae). *Mol. Phylogenet. Evol.* 65: 926–939.
- 601 Goubert, C., L. Modolo, C. Vieira, C. ValienteMoro, P. Mavingui *et al.*, 2015 De novo assembly  
602 and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE  
603 from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes*  
604 *aegypti*). *Genome Biol. Evol.* 7: 1192–1205.
- 605 Harpur, B. A., A. Dey, J. R. Albert, S. Patel, H. M. Hines *et al.*, 2017 Queens and workers  
606 contribute differently to adaptive evolution in bumble bees and honey bees. *Genome Biol.*  
607 *Evol.* 9: 2395–2402.
- 608 Harpur, B. A., C. F. Kent, D. Molodtsova, J. M. Lebon, A. S. Alqarni *et al.*, 2014 Population  
609 genomics of the honey bee reveals strong signatures of positive selection on worker traits.  
610 *Proc Natl Acad Sci U S A* 111: 2614–2619.

- 611 James, R. R., 2011 Bee importation, bee price data, and chalk-brood, in *Western Alfalfa Seed*  
612 *Grower Association Winter Seed Conference*, Las Vegas, NV.
- 613 Johansen, C. A., D. F. Mayer, and J. D. Eves, 1978 *Biology and management of the alkali bee,*  
614 *Nomia melanderi Cockrell (Hymenoptera: Halictidae)*. Washington State Entomology.
- 615 Johnson, A. D., R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell *et al.*, 2008 SNAP:  
616 A web-based tool for identification and annotation of proxy SNPs using HapMap.  
617 *Bioinformatics* 24: 2938–2939.
- 618 Joshi, N. ., and J. . Fass, 2011 Sickle: A sliding-window, adaptive, quality-based trimming tools  
619 for FastQ files.
- 620 Kapheim, K. M., 2017 Nutritional, endocrine, and social influences on reproductive physiology  
621 at the origins of social behavior. *Curr. Opin. Insect Sci.* 22: 62–70.
- 622 Kapheim, K. M., and M. M. Johnson, 2017a Juvenile hormone, but not nutrition or social cues,  
623 affects reproductive maturation in solitary alkali bees (*Nomia melanderi*). *J. Exp. Biol.*  
624 *jeb.162255*.
- 625 Kapheim, K. M., and M. M. Johnson, 2017b Support for the reproductive ground plan hypothesis  
626 in a solitary bee: Links between sucrose response and reproductive status. *Proc. R. Soc. B*  
627 *Biol. Sci.* 284: 20162406.
- 628 Kapheim, K. M., H. Pan, C. Li, S. L. Salzberg, D. Puiu *et al.*, 2015 Social evolution. Genomic  
629 signatures of evolutionary transitions from solitary to group living. *Science* 348: 1139–  
630 1143.
- 631 Kocher, S. D., C. Li, W. Yang, H. Tan, S. V Yi *et al.*, 2013 The draft genome of a socially  
632 polymorphic halictid bee, *Lasioglossum albipes*. *Genome Biol.* 14: R142.
- 633 Li, H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

- 634 ARXIV 1303.3997:
- 635 Liu, H., Y. Jia, X. Sun, D. Tian, L. D. Hurst *et al.*, 2017 Direct determination of the mutation rate  
636 in the bumblebee reveals evidence for weak recombination-associated mutation and an  
637 approximate rate constancy in insects. *Mol. Biol. Evol.* 34: 119–130.
- 638 Löytynoja, A., 2014 Phylogeny-aware alignment with PRANK, pp. 155–170 in *Multiple*  
639 *Sequence Alignment Methods. Methods in Molecular Biology (Methods and Protocols)*,  
640 edited by D. Russell. Humana Press, Totowa, NJ.
- 641 Penn, O., E. Privman, H. Ashkenazy, G. Landan, D. Graur *et al.*, 2010 GUIDANCE: a web  
642 server for assessing alignment confidence scores. *Nucleic Acids Res.* 38: W23-8.
- 643 Picard. <http://picard.sourceforge.net/>. Accessed 12 January 2016.
- 644 Price, A. L., N. C. Jones, and P. A. Pevzner, 2005 De novo identification of repeat families in  
645 large genomes. *Bioinformatics* 21: i351–i358.
- 646 Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder *et al.*, 2005 InterProScan: protein  
647 domains identifier. *Nucleic Acids Res* 33: W116-20.
- 648 Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing  
649 genomic features. *Bioinformatics* 26: 841–842.
- 650 Rehan, S. M., and A. L. Toth, 2015 Climbing the social ladder: The molecular evolution of  
651 sociality. *Trends Ecol. Evol.* 30: 426–433.
- 652 Robinson, G. E., C. M. Grozinger, and C. W. Whitfield, 2005 Sociogenomics: Social life in  
653 molecular terms. *Nat. Rev. Genet.* 6: 257–270.
- 654 Romiguier, J., J. Lourenco, P. Gayral, N. Faivre, L. A. Weinert *et al.*, 2014 Population genomics  
655 of eusocial insects: the costs of a vertebrate-like effective population size. *J. Evol. Biol.* 27:  
656 593–603.

- 657 Sadd, B. M., S. M. Barribeau, G. Bloch, D. C. De Graaf, P. Dearden *et al.*, 2015 The genomes of  
658 two key bumblebee species with primitive eusocial organization. *Genome Biol.* 16: 76.
- 659 Sinha, S., Y. Liang, and E. Siggia, 2006 Stubb: a program for discovery and analysis of cis -  
660 regulatory modules. *Nucleic Acids Res.* 34: W555–W559.
- 661 Smit AFA, Hubley R, Green P: RepeatMasker. Available at: <http://www.repeatmasker.org>.  
662 [Accessed 9 April 2013].
- 663 Stamatakis, A., 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of  
664 large phylogenies. *Bioinformatics* 30: 1312–3.
- 665 Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack *et al.*, 2006 AUGUSTUS: A b initio  
666 prediction of alternative transcripts. *Nucleic Acids Res.* 34:.
- 667 Stephen, W. P., 2003 Solitary bees in North American agriculture: a perspective, pp. 41–66 in  
668 *For Nonnative Crops, Whence Pollinators of the Future?*, edited by K. Strickler and J. H.  
669 Cane. Entomol. Soc. Am., Lanham, MD.
- 670 Stolle, E., R. Pracana, P. Howard, C. I. Paris, S. J. Brown *et al.*, 2018 Degenerative expansion of  
671 a young supergene. bioRxiv 326645.
- 672 Suyama, M., D. Torrents, and P. Bork, 2006 PAL2NAL: robust conversion of protein sequence  
673 alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34: W609-12.
- 674 Tajima, F., 1989 The effect of change in population size on DNA polymorphism. *Genetics* 123:.
- 675 Team, R. C., 2016 R: A language and environment for statistical computing.
- 676 Terhorst, J., J. A. Kamm, and Y. S. Song, 2017 Robust and scalable inference of population  
677 history from hundreds of unphased whole genomes. *Nat. Genet.* 49: 303–309.
- 678 Trapnell, C., L. Pachter, and S. L. Salzberg, 2009 TopHat: Discovering splice junctions with  
679 RNA-Seq. *Bioinformatics* 25: 1105–1111.

- 680 Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim *et al.*, 2012 Differential gene and transcript  
681 expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7:  
682 562–578.
- 683 U.S. Department of Agriculture, N. A. S. S., 2014 Census of Agriculture Summary and State  
684 Data 2012: U.S. Government Printing Office.
- 685 Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis *et al.*, 2018 BUSCO  
686 Applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol.*  
687 *Evol.* 35: 543–548.
- 688 Wcislo, W. T., and M. S. Engel, 1996 Social behavior and nest architecture of nomiine bees  
689 (Hymenoptera: Halictidae; Nomiinae). *J. Kansas Entomol. Soc.* 69: 158–167.
- 690 Wcislo, W. T., and J. H. Fewell, 2017 Sociality in bees, pp. 50–83 in *Comparative Social*  
691 *Evolution*, edited by D. R. Rubenstein and P. Abbot. Cambridge University Press,  
692 Cambridge.
- 693 Wilson, E. O., 1971 *The Insect Societies*. Harvard University Press, Cambridge, Massachusetts.
- 694 Xu, Z., and H. Wang, 2007 LTR\_FINDER: an efficient tool for the prediction of full-length LTR  
695 retrotransposons. *Nucleic Acids Res.* 35: W265–W268.
- 696 Yang, Z., 2007 PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:  
697 1586–1591.
- 698 Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders *et al.*, 2010 Common SNPs  
699 explain a large proportion of heritability for human height. *Nat. Genet.* 42: 565–569.
- 700 Zdobnov, E. M., and R. Apweiler, 2001 InterProScan--an integration platform for the signature-  
701 recognition methods in InterPro. *Bioinformatics* 17: 847–848.
- 702 Zdobnov, E. M., F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simão *et al.*, 2017



703 OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant,  
704 archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 45: D744–D749.

705 Zhu, L. J., R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh *et al.*, 2011

706 FlyFactorSurvey: A database of *Drosophila* transcription factor binding specificities  
707 determined using the bacterial one-hybrid system. *Nucleic Acids Res.* 39: 111–117.

708

## 709 **ACKNOWLEDGEMENTS**

710 We are grateful to M. Ingham, M. Buckley, and M. Wagoner for allowing us to collect from their  
711 bee beds. J. Dodd and Forage Genetics International provided lab space while in the field.

712 Sequencing was performed by the Roy J. Carver Biotechnology Center at University of Illinois  
713 at Urbana-Champaign (UIUC). Computational support was provided by University of Utah

714 Center for High Performance Computing and UIUC CNRG/Biocluster. J. Johnson (Life Sciences  
715 Studios) created the illustration in Fig. 1. Funding was provided by Utah Agricultural

716 Experiment Station project 1297 (KMK) and grants from the USDA-ARS Alfalfa Pollinator  
717 Research Initiative (KMK) and USDA-NIFA grant # 2018-67014-27542 (KMK). Additional

718 funding was provided by a Smithsonian Institution Competitive Grants Program for

719 Biogenomics (WTW, KMK, BMJ), Swiss National Science Foundation grant PP00P3\_170664

720 (RMW), and general research funds from the Smithsonian Tropical Research Institute (WTW).

721

## 722 **SUPPLEMENTARY FILES**

723 **Table S1. Selected species for orthology analysis.** All protein sets were collated from OrthoDB

724 v9.1, except the new genome assembly presented here (*Nomia melanderi*) and the unpublished

725 *Megalopta genalis* genome assembly (Kapheim *et al.*, unpublished; BioProject PRJNA494872).

726

727 **Table S2. Comparison of the eight selected Apoidea genomes used for orthology and**  
728 **phylogenomics analyses, as well as BUSCO assessments.**

729

730 **Table S3. BUSCO genome and gene set assessments of selected Hymenoptera.** Complete  
731 BUSCOs (C); Complete and single-copy BUSCOs (S); Complete and duplicated BUSCOs (D);  
732 Fragmented BUSCOs (F); Missing BUSCOs (M); Total BUSCO groups searched (n) for Insecta  
733 (Ins.) and Hymenoptera (Hym.) assessment sets.

734

735 **Table S4. Genes under selection.** This is a list of *N. melanderi* genes found to undergoing  
736 positive selection via population genetic analyses. Orthogroup IDs and orthogroup ages are also  
737 included.

738

739 **Table S5. GO enrichment for genes under selection.** This is the GOstats output for enrichment  
740 tests of the set of genes under positive selection in *N. melanderi*.

741

742 **Table S6. Transcription factor motif enrichments among genes under positive selection in**  
743 ***N. melanderi*.**

744

745 **Table S7. Model likelihood scores and predicted omega values from PAML.** Omega values  
746 and model likelihood results from PAML analysis of 6676 orthogroups. Omega values are given  
747 for *N. melanderi* for best-fitting model(s), as well as additional foreground and background  
748 branches. Raw and FDR corrected p-values for each model likelihood test are also provided.

749 Orthogroups with  $dS > 2$  in at least one branch in the best-fitting model were filtered from the  
750 results presented in the main text.

751

752 **Table S8. GO enrichment for PAML results.** This is the GOstats output for functional  
753 enrichment tests on the set of genes identified as evolving faster or slower in *N. melanderi*, as  
754 compared to other species in Halictidae.

755

756 **File S1. VCF file.** File containing genetic variant information for the 18 females included in the  
757 population genetic analysis.

758

759 **File S2. Collection of normalized TF motif binding score matrices.** The ten files in this  
760 archive are matrices of gene-level motif binding scores for *N. melanderi* for each combination of  
761 the two different normalization procedures ("rank" and "gc") and the five different regulatory  
762 region definitions. The rows are genes, the columns motifs, and the scores are the length-  
763 adjusted normalized scores which range from 0 (best) to 1 (worst). Scores of "2" signal a missing  
764 motif score, likely indicating that the regulatory region was small and masked by tandem repeats.  
765 More details and datasets for additional bee species are available at

766 <http://veda.cs.uiuc.edu/beeMotifScores/>.

767

768 **File S3. Repetitive Elements in the genome of *Nomia melanderi* (Nme).** This contains  
769 supplementary information on the repetitive and transposable elements in Nme, including  
770 additional figures (S3.Nme.Repeats.report.pdf), tables (Nme.basepairs.by.type.txt: Nme  
771 basepairs per repeat type; Nme.TE.groups.counts.txt: Nme counts per TE group;

772 Nme.TE.elements.counts.txt: Nme counts per TE elements, Nme.TE.RM.annotation.report.txt:  
773 Nme repeats annotated with RepeatMasker), and annotation files (Nme.TE.fa: Nme repetitive  
774 elements fasta sequences; Nme.TE.gff: Nme sequence assembly annotation of repetitive  
775 elements (gff)).

776

777 **File S4. Predicted SNP effects.** This file has two tabs describing the predicted effects of each  
778 filtered SNP. The ‘functionalPredictions’ tab indicates the number of SNPs found within each  
779 protein-coding gene, and its position/effect relative to the gene structure. The tab  
780 ‘predictedEffects’ describes the location (Scaffold, Position) of each filtered SNP associated with  
781 a protein coding gene. The reference and alternate alleles are provided, along with the gene name  
782 it is associated with.

783

784 **File S5. Gene annotation file (GFF) for *N. melanderi*.**