The nuclear and mitochondrial genomes of the facultatively eusocial orchid bee *Euglossa dilemma.*

Authors:

Philipp Brand[*,†], Nicholas Saleh[*,†], Hailin Pan[‡], Cai Li[‡], Karen M. Kapheim[§], Santiago R. Ramírez[*]

Affiliations:

* Department for Evolution and Ecology, Center for Population Biology, University of California, Davis, California 95616

† Graduate Group in Population Biology, University of California, Davis, California 95616

‡ China National Genebank, BGI-Shenzhen, Shenzhen, 518083, China

§ Department of Biology, Utah State University, Logan, Utah 84322

Data availability

The *E. dilemma* genome assembly *Edil_v1.0*, the annotation, and the original gene set *Edil_OGS_v1.0* are available for download via NCBI [XXX], Beebase [XXX], the i5k NAL workspace [xxx], and the Ramirez Lab website [URL]. The raw reads are available at the NCBI Sequence Read Archive [XXX]. The published raw transcriptome sequence reads are available at the NCBI Sequence Read Archive [SRA: SRX765918].

Running Title:

The genomes of the orchid bee *Euglossa dilemma*

Key Words:

whole-genome assembly; corbiculate bee; orchid bee; invasive species; mitochondrial genome

Corresponding Author:

Philipp Brand
Department of Evolution & Ecology
University of California, Davis
One Shields Ave
Davis, CA 95616

Office: (530) 752-7614
pbrand@ucdavis.edu

Abstract

Bees provide indispensable pollination services to both agricultural crops and wild plant populations, and several species of bees have become important models for the study of learning and memory, plant-insect interactions and social behavior. Orchid bees (Apidae: Euglossini) are especially important to the fields of pollination ecology, evolution, and species conservation. Here we report the nuclear and mitochondrial genome sequences of the orchid bee *Euglossa dilemma*. *Euglossa dilemma* was selected because it is widely distributed, highly abundant, and it was recently naturalized in the southeastern United States. We provide a high-quality assembly of the 3.3 giga-base genome, and an official gene set of 15,904 gene annotations. We find high conservation of gene synteny with the closely related honey bee. This genomic resource represents the first draft genome of the orchid bee genus *Euglossa,* and the first draft orchid bee mitochondrial genome, thus representing a valuable resource to the research community.

Introduction

Bees (Apoidea) are important models for the study of learning and memory (Menzel and Muller 1996), plant-insect interactions (Doetterl and Vereecken 2010) and the evolution of social behavior (Nowak *et al.* 2010; Woodard *et al.* 2011; Kapheim *et al.* 2015). Among the >20,000 bee species worldwide, lineages have evolved varied degrees of specialization on floral resources such as pollen, resins, and oils (Wcislo and Cane 2003; Michener 2007; Litman *et al.* 2011). These relationships are wide-ranging and have substantial impact on the health and function of natural and agricultural systems (Klein *et al.* 2007). Furthermore, several transitions from an ancestral solitary to a derived eusocial behavior have occurred within bees (Danforth 2002; Cardinal and Danforth 2011; Branstetter *et al.* 2017). Thus, bees provide unique opportunities to investigate the genetic underpinnings of multiple major ecological and evolutionary transitions. The repeated evolution of different behavioral phenotypes in bees, including foraging and social behavior, provides a natural experiment that allows for the determination of general as well as species-specific molecular genomic changes underlying phenotypic transitions. In order to capitalize on this potential, whole-genome sequences of a divergent array of bee species with different life histories are needed (Kapheim *et al.* 2015).

Orchid bees (Apidae; Euglossini) are among the most important pollinators of thousands of diverse neotropical plant species (Ramírez *et al.* 2002). While female orchid bees collect nectar, pollen and resin for nest construction and provisioning, male bees collect perfume compounds from floral and non-floral sources (Vogel 1966; Whitten *et al.* 1993; Eltz *et al.* 1999; Roubik and Hanson 2004). These volatile compounds are used to concoct a species-specific perfume blend that is subsequently used during courtship, presumably to attract conspecific females. This unique male scent-collecting behavior has recently been examined in a broad array of molecular ecological and evolutionary studies, focusing on phenotypic evolution,

chemical communication, plant-insect mutualisms, and speciation (Eltz *et al.* 2008; 2011; Ramírez *et al.* 2011; Brand *et al.* 2015).

While most of the approximately 220 species of orchid bees appear to be solitary, several species have transitioned to living in coordinated social groups (Garófalo 1985; Pech *et al.* 2008; Augusto and Garófalo 2009). Female *Euglossa dilemma* individuals, for example, can either form solitary nests and provision their own brood cells, or live in small groups where daughters remain in their natal nest and help their mother rear her offspring, instead of dispersing to found their own nest. The social *E. dilemma* nests (similar to the closely related *E. viridissima;* Pech *et al.* 2008) exhibit true division of labor, with subordinate daughters foraging for resources and the reproductively dominant mothers laying eggs (Saleh & Ramirez pers. obs.). This facultative eusocial behavior represents an early stage in social evolution and makes *E. dilemma* well-suited for studying the genomic changes underlying the transition from solitary to eusocial behavior. While other facultative eusocial species evolved throughout the bee lineage, orchid bees have a unique taxonomic position (Cardinal and Danforth 2011). Orchid bees are part of the corbiculate bees, together with the honey bees, bumblebees and stingless bees, three obligately eusocial bee lineages (Figure 1a). As the sister group to all other corbiculate bee lineages (Romiguier *et al.* 2015; Peters *et al.* 2017; Branstetter *et al.* 2017), orchid bees may provide key insights into the early stages of eusociality and the possible evolutionary trajectories that led to the obligate eusocial behavior exhibited by honey bees.

Here we present the draft genome of the orchid bee species *Euglossa dilemma*. Using a combined paired-end and mate-pair library sequencing approach, we assembled 18% of the predicted 3.2Gb genome, and annotated a high-quality gene set including 15,904 genes. In addition, we reconstructed three quarters of the mitochondrial genome with the help of transcriptome data, representing the first orchid bee mitogenome. These genomic resources will facilitate the genetic study of outstanding ecological and evolutionary questions, such as the evolution of resource preferences and the evolution of eusociality. Moreover, it provides an important genomic resource for an endangered group of crucial neotropical pollinators, that are of specific concern for conservation biologists (Zimmermann *et al.* 2011; Suni and Brosi 2012; Suni 2016; Soro *et al.* 2016).

Materials and Methods

Genome sequencing and assembly

Nuclear Genome. Sequencing of the *E. dilemma* genome was based on six haploid male individuals collected at Fern Forest Nature Center in Broward County, FL in February 2011. This population was chosen due to its low nucleotide diversity resulting from a bottleneck during a single introduction to Southern Florida about 15 years ago (Skov and Wiley 2005; Pemberton and Wheeler 2006; Zimmermann *et al.* 2011). DNA was extracted from each bee independently and used for the

184 construction of four paired-end (two 170bp and 500bp libraries, respectively) and
185 four mate-pair (two 2kb and 5kb libraries, respectively) sequencing libraries. Next,
186 the paired-end libraries were sequenced in 90 cycles and the mate-pair libraries for
187 49 cycles on an Illumina HiSeq2000. The resulting sequence data was run through
188 fastuniq v1.1 (Xu *et al.* 2012) to remove PCR duplicates and quality trimmed using
189 trim_galore v0.3.7 (Babraham Bioinformatics). Subsequently, reads were used for *de*
190 *novo* assembly with ALLPATHS-LG v51750 (Gnerre *et al.* 2011) and Soap-denovo2
191 (Luo *et al.* 2012) with varying settings. Gaps within scaffolds were closed using
192 GapCloser v1.12 (Luo *et al.* 2012) for each assembly. ALLPATHS-LG with default
193 settings resulted in the highest-quality assembly, based on assessments of
194 annotation completeness (see below). This assembly (*E. dilemma* genome assembly
195 v1.0) was used for all subsequent analyses. All other assemblies were excluded from
196 analysis, but are available upon request.
197
198 The pre-processed reads were used for k-mer based genome size estimates. We
199 used ALLPATHS-LG to produce and analyze the k-mer frequency spectrum (k=25).
200 Genome size was estimated on the basis of the consecutive length of all reads
201 divided by the overall sequencing depth as $(N \times (L - K + 1) - B)/D = G$, where N is
202 the total number of reads, L is the single-read length, K is the k-mer length, B is the
203 total count of low-frequency (frequency ≤3) k-mers that are most likely due to
204 sequencing errors, D is the k-mer depth estimated from the k-mer frequency
205 spectrum, and G is the genome size. In addition, we used the ALLPATHS-LG k-mer
206 frequency spectrum to predict the repetitive fraction of the genome.
207
208 The quality of the genome assembly was assessed using standard N statistics, and
209 assembly completeness as measured by the CEGMA v2.5 (Parra *et al.* 2007) and
210 BUSCO v1.1 (Simão *et al.* 2015) pipelines. CEGMA was run in default mode, whereas
211 BUSCO was run with the arthropoda_odb9 OrthoDB database (Zdobnov *et al.* 2017)
212 in genome mode.
213
214 We estimated the mean per-base genome coverage on the basis of the pre-processed
215 reads and the estimated genome size as $\frac{\sum_{i=1}^{4}(R_i * L_i)}{G} = C$, where R is the number of
216 reads and L the mean read length of sequence library i, G is the estimated genome
217 size and C the resulting per-base coverage.
218
219 *Mitogenome.* Initial attempts to reconstruct the mitochondrial genome from our
220 whole-genome shotgun sequencing reads were only partially successful, due to high
221 sequence variability of sequencing reads with similarity to mitochondrial loci (data
222 not shown). In addition, we have observed that the amplification of mitochondrial
223 DNA in standard polymerase chain reactions (PCR) leads to a high level of
224 polymorphic sites in *E. dilemma* and other orchid bees (Brand & Ramírez pers. obs.).
225 Together, this suggests the presence of nuclear copies of the mitochondrial genome
226 (NUMTs) that interfere with the assembly process and PCR amplification. Therefore,
227 we used available *E. dilemma* transcriptome assemblies in order to reconstruct the
228 mitochondrial genome from cDNA (Brand *et al.* 2015). In order to find

229 mitochondrial genes in the transcriptome assembly of Brand et al. 2015, we used
230 blastx with the honey bee mitochondrial genome as query (Crozier and Crozier
231 1993) and an E-value cutoff of 10E-12 (Altschul *et al.* 1990; Camacho *et al.* 2009).
232 The contigs and scaffolds that were detected with this approach were annotated
233 following Dietz *et al.* 2015. Briefly, we performed tblastn and tblastx searches with
234 protein coding genes and rRNA genes of the honey bee mitochondrial genome,
235 respectively. All hits were used for manual gene annotation using Geneious v8.0.5
236 (Biomatters Ltd. 2012). Since the recovered mitochondrial mRNA scaffolds
237 contained more than one gene, we searched and annotated intergenic tRNAs using
238 ARWEN 1.2.3 (Laslett and Canbäck 2008) and tRNAscan-SE 1.21 (Lowe and Eddy
239 1997).
240
241 **Genome annotation**
242
243 *Gene annotation.* Genes were annotated based on sequence homology and *de novo*
244 gene predictions. The homology approach was based on the recently updated high-
245 quality official gene set of the honey bee (OGS v3.2; Elsik *et al.* 2014). All honey bee
246 original gene set (OGS) proteins were used in initial tblastn searches against all *E.*
247 *dilemma* scaffolds with an E-value cutoff of 10E-4. Proteins with a hit covering ≥
248 50% of the honeybee protein query were selected for accurate exon-intron
249 boundary prediction for each scaffold using exonerate v2.42.1 (Slater and Birney
250 2005) with the minimum fraction of the possible optimal similarity per query set to
251 35%. In a second round, genes not annotated under the previous settings were
252 rerun with minimum similarity set to 15%. In the case of multiple annotations, we
253 discarded all but one annotation with the best hit to the honeybee OGS (based on
254 completeness and similarity). This approach proved feasible due to the close
255 relatedness of *E. dilemma* and the honey bee. For *de novo* gene prediction we used
256 Augustus (Stanke *et al.* 2008) and SNAP (Korf 2004) trained on the honey bee, with
257 the *E. dilemma* genome masked for repetitive regions (See below) as input. Only
258 genes predicted by both programs were taken into account. Gene predictions with
259 ≥85% sequence similarity to each other were discarded, to prevent the inclusion of
260 putative unmasked transposable element derived genes in the official gene set. *De*
261 *novo* predictions were added to the *E. dilemma* OGS if not annotated by the
262 homology-based approach.
263
264 *Repetitive element annotation.* Repetitive elements including tandem repeats,
265 nuclear copies of the mitochondrial genome (NUMTs), and transposable elements
266 (TEs) were annotated using multiple methods.
267
268 *Tandem repeats.* We searched for micro- and mini-satellites (1–6 bp and 7–1000 bp
269 motif length, respectively) in all scaffolds using Phobos 3.3.12 (Mayer 2010). We
270 performed two independent runs for each class of tandem repeats with Phobos
271 parameter settings following Leese et al. 2012 (gap score and mismatch score set to
272 -4 and a minimum repeat score of 12; Leese *et al.* 2012).
273

274 *NUMTs*. We annotated NUMTs using blastn runs with the partial mitochondrial
275 genome (see above) as query and an E-value cutoff of 10E-4 as used in NUMT
276 analyses of other insect genomes (Pamilo *et al.* 2007). This approach allowed us to
277 find NUMTs with medium to high similarity to the actual transcriptome-based
278 mitochondrial genome.
279
280 *TEs*. In order to annotate TEs, we used Repeatmasker v4.0.5 (Smit et al. 2016) with
281 Crossmatch v. 0.990329 as search engine in sensitive mode. The Repbase
282 invertebrate database v21.12 (Jurka 2000; Bao *et al.* 2015) was used as TE
283 reference, due to the lack of a bee-specific database.
284
285 **Genome structure**
286
287 To analyze genome structure, we compared the genome wide gene synteny of *E.*
288 *dilemma* and the honey bee. We used the genomic locations of homologous genes (as
289 determined above) of the honey bee and *E. dilemma* scaffolds of at least 100kb
290 length to build haplotype blocks with a minimum length of 1kb. Haplotype blocks
291 included the entire gene span as well as intergenic regions whenever two or more
292 adjacent genes were homologous in both species. We discarded gene annotations
293 from downstream analysis that were recovered as homologous to multiple genomic
294 locations in either species. Furthermore, we excluded *E. dilemma* genes that were
295 recovered as homologous to honey bee scaffolds belonging to unknown linkage
296 groups.
297
298 **Data availability**
299
300 The *E. dilemma* genome assembly *Edil_v1.0*, the annotation, and the original gene set
301 *Edil_OGSv1.0* are available for download via NCBI [XXX], Beebase [XXX] (Elsik *et al.*
302 2016), the i5k NAL workspace [xxx] (i5K Consortium 2013), and the Ramirez Lab
303 website [URL]. The raw reads are available at the NCBI Sequence Read Archive
304 [XXX]. The published raw transcriptome sequence reads are available at the NCBI
305 Sequence Read Archive [SRA: SRX765918] (Brand *et al.* 2015).

306
307 **Results and Discussion**
308
309 **Whole-genome assembly**
310
311 The *E. dilemma* genome assembly resulted in 22,698 scaffolds with an N50 scaffold
312 length of 144Kb and a total length of 588Mb (Table 1). This represents 18% of the k-
313 mer based estimated genome size of 3.2Gb. Of all sequence reads, 68% aligned to
314 the genome assembly, of which 56% aligned more than once. Further, the k-mer
315 frequency spectrum based on all sequencing reads was strongly positively skewed
316 indicating the presence of highly repetitive sequences in the read set (Figure 1b).
317 Based on the k-mer frequency spectrum, 87.7% of the genome was estimated to be

318 repetitive. This suggests that the genome of *E. dilemma* consists largely of highly
319 repetitive sequences, explaining the low consecutive assembly length and the high
320 assembly fragmentation. The mean per-base coverage was estimated to be
321 comparatively low in comparison to previous bee genome assemblies, with 19.7x
322 based on the pre-processed reads and estimated genome size (Kocher *et al.* 2013;
323 Kapheim *et al.* 2015). Total genomic GC content was 39.9%, and thus similar to
324 previously sequenced bee genomes ranging between 32.7% and 41.5% (Table 1)
325 (Kocher *et al.* 2013; Elsik *et al.* 2014; Kapheim *et al.* 2015).
326
327 Despite the fragmentation of the genome assembly representing less than 20% of
328 the estimated genome size, CEGMA analysis revealed complete assemblies of 231
329 out of 248 core eukaryotic genes (93.2% completeness). Similarly, BUSCO analysis
330 revealed that 1007 out of 1066 highly conserved arthropod genes were completely
331 assembled (94.4% completeness). Our gene prediction approach generated a
332 comprehensive official gene set including 15,904 protein-coding genes (Table 1). Of
333 these gene models, 11,139 were derived from homology-based predictions,
334 representing 73% of the 15,314 honey bee genes used for annotation. These
335 annotations are well within or exceeding previous bee genome assemblies, and are
336 similar to those reported for the other orchid bee genome available (Table 1)
337 (Kocher *et al.* 2013; Elsik *et al.* 2014; Park *et al.* 2015; Sadd *et al.* 2015; Kapheim *et*
338 *al.* 2015).
339
340 The CEGMA and BUSCO analysis and the gene annotation results suggest that the
341 gene-coding fraction of the *E. dilemma* genome was properly assembled, despite the
342 large estimated genome size and comparatively low per-base sequencing coverage.
343 Fragmentation of the assembly is thus likely to be primarily the result of highly
344 repetitive genomic elements, and less the result of low coverage. Overall, the results
345 suggest that our approach was sufficient to produce a high quality official gene set.
346 Due to the chosen homology-based approach, the majority of annotated genes in the
347 official gene set has known homology to honey bee genes (Table S1), which will
348 greatly facilitate genome-wide expression studies including gene ontology analyses
349 and comparative gene set analyses among insects.
350
351 **Mitochondrial Genome assembly**
352 The recently published transcriptome assembly used for the reconstruction of the
353 mitochondrial genome contained four scaffolds between 1,222bp and 4,188bp long
354 with a total consecutive length of 11,128bp (Figure 2). This corresponds to about
355 75% of the estimated length of the mitochondrial genome, based on other
356 corbiculate bee species (Crozier and Crozier 1993; Cha *et al.* 2007). The *E. dilemma*
357 mitogenome fragments contained 5 out of 22 tRNAs, 11 out of 13 protein coding
358 genes of which two were only partially recovered, and the 16S rRNA gene. Within
359 scaffolds all genes showed the known hymenopteran gene order and orientation,
360 while the orientation of the 5 tRNAs detected was identical to those in the honey bee
361 (Crozier and Crozier 1993; Cha *et al.* 2007). Attempts to complete the mitochondrial
362 genome using the nuclear genome assembly yielded no improvement of the
363 assembly (data not shown).

364
365 The high success in mitochondrial gene reconstruction is likely due to the nature of
366 the analyzed transcriptome data. Short intergenic regions as well as polycistronic
367 mitochondrial mRNA likely lead to the assembly of multiple genes into single
368 scaffolds. The A-T rich region is completely missing as well as the ND2 and 12S
369 rRNA genes flanking the region in insect mitogenomes. This unrecovered region also
370 contains a high number of tRNAs in the honeybee, which could explain the low
371 number of recovered tRNAs in *E. dilemma*. While the partial mitochondrial genome
372 assembly is only 75% complete, it represents the first mitogenome for the group of
373 orchid bees and will thus be a valuable resource for future phylogenetic analyses
374 within the lineage and between more distantly related bee taxa.
375
376 **Repetitive elements**
377
378 *Tandem Repeats*. We detected 76,001 microsatellite loci with a consecutive length of
379 2,291,067 bp. Minisatellites with motif lengths from 7bp to 1000bp were less
380 numerous in the genome (67,323 loci), totaling 13,343,515bp. Accordingly, tandem
381 repeats represent 3.86% of the genome assembly, suggesting that they contribute
382 only a small proportion to the overall genome size (Figure 1c).
383
384 *NUMTs*. We detected fragments with similarity to the draft mitochondrial genome
385 on 129 scaffolds totaling a length of 150,670 bp. The fragments had a mean length of
386 764.8 bp and a mean similarity of 91.5% to the mitogenome. This suggests that
387 these fragments are not derived from the mitochondrial genome and represent
388 actual NUMTs. A total of 39 scaffolds carried multiple fragments with high similarity
389 to the mitogenome with a concatenated length of up to 6566bp, suggesting that
390 respective NUMTs might have originated from larger fragments of the mitogenome.
391 In total, only 0.04% of the whole-genome assembly had hits to the mitogenome
392 (Figure 1c). This is likely an underestimate, due to the incompleteness of the
393 reconstructed mitochondrial genome. Nevertheless, NUMTs likely represent only a
394 small fraction of the whole nuclear *E. dilemma* genome. Previous analyses have
395 shown a high density of NUMTs in the honey bee in comparison to other insect
396 genomes totaling about 0.1% of the overall genome size (Pamilo *et al.* 2007).
397 Accordingly, given the NUMT content detected in *E. dilemma*, it is possible that a
398 comparatively high NUMT density is a common feature of corbiculate bee genomes.
399
400 *TEs*. In our RepeatMasker analysis we detected 47,553 interspersed repeats with a
401 cumulative length of 7,747,824bp. This corresponds to 1.82% of the genome
402 assembly (Figure 1c). This low number is surprising, since large genome sizes as in
403 *E. dilemma* have been mainly associated with TE activity and content in genomes
404 from unicellular eukaryotes to complex multicellular organisms including plants,
405 invertebrates and vertebrates (Kidwell 2002). However, TEs are fast evolving and
406 highly specific to their host lineages, which leads to large underestimates of genomic
407 TE content in previously unstudied lineages (Chalopin *et al.* 2015; Platt *et al.* 2016).
408 The only bee repeat content included in the Repbase database used for TE
409 annotation is the honey bee, a species with a comparatively small genome (0.23Gb)

410 and low TE diversity and content (Weinstock *et al.* 2006; Kapheim *et al.* 2015). Thus,
411 the low percentage of TEs detected in the *E. dilemma* genome is very likely an
412 underestimate of the actual density. *De novo* TE reconstructions using the genomic
413 resources presented here should be performed in the future to provide a better
414 estimate of the actual TE density in the *E. dilemma* genome.
415
416 **Genome structure**
417
418 Of the 22,698 *E. dilemma* scaffolds, 580 were at least 100kb in length and used for
419 synteny analysis with the honey bee genome. A total of 356 of these scaffolds
420 carried at least one gene annotation with known homology to the honey bee, and
421 329 of these *E. dilemma* scaffolds were homologous to honey bee scaffolds with
422 known linkage group (LG) association (Table S1). Of these scaffolds, 272 (83%)
423 showed ≥95% syntenic homology to a single honey bee LG (Figure 1d). Overall, the
424 detected syntenic linkage blocks cover 222MB of scaffold length with homology to
425 the honeybee, representing 85% of the 329 filtered scaffolds. Mean syntenic block
426 length is 206,033bp for *E. dilemma* (min: 1017bp; max: 4,458,123bp; median:
427 47,154bp) and 143,221.4bp for the honey bee (min: 1029bp; max: 2,205,566bp;
428 median: 41,726bp). This suggests that the genomic architecture is very similar
429 between *E. dilemma* and the honey bee, representing a high level of conservation
430 during the 80 million years since the two lineages diverged. Further, our results
431 support a recent comparative analysis of the honey bee and the bumblebee
432 genomes, which revealed high conservation of genomic synteny (Stolle *et al.* 2011).
433 Together, these results support a general pattern of surprisingly slow genome
434 evolution in gene coding regions in corbiculate bees, independent of the fraction of
435 repetitive genome content.
436
437 **Conclusion**
438
439 The genome assembly of the orchid bee *E. dilemma* that we present here is of high
440 quality, despite its large genome size (estimated to be 3.3Gb). The 15,904 gene
441 annotations provide a comprehensive set of genes with known homology to the
442 honey bee, facilitating future gene ontology and functional genomic analyses. While
443 we were unable to annotate the mostly repetitive majority of the genome assembly
444 with our approach, the provided sequence reads will be useful for future analyses of
445 repetitive genetic elements in the genome. The nuclear and mitochondrial draft
446 genomes represent a valuable genomic resource for the community of bee
447 geneticists. This genomic resource will likely prove valuable in genetic and
448 functional genomic analyses dealing with the ecology, evolution, and conservation of
449 orchid bees. Furthermore, the genome of the facultatively eusocial *E. dilemma* will
450 be helpful in the study of the evolution of eusociality, due to its taxonomic
451 placement as the sister lineage to the three obligately eusocial corbiculate bee tribes
452 including stingless bees, bumblebees, and honey bees.
453
454 **Acknowledgements**
455 We would like to thank Gene E. Robinson, Guojie Zhang, and the BGI for financial

support for genome sequencing. P.B. was supported by a fellowship of the German academic exchange service (Deutscher Akademischer Austauschdienst, DAAD) for parts of the project.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Augusto, S. C., and C. A. Garófalo, 2009 Bionomics and sociological aspects of Euglossa fimbriata (Apidae, Euglossini). Genet. Mol. Res. 8: 525–538.

Bao, W., K. K. Kojima, and O. Kohany, 2015 Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 2015 6:1 6: 11.

Brand, P., S. R. Ramírez, F. Leese, J. J. Quezada-Euan, R. Tollrian *et al.*, 2015 Rapid evolution of chemosensory receptor genes in a pair of sibling species of orchid bees (Apidae: Euglossini). BMC Evol. Biol. 15: 176.

Branstetter, M. G., B. N. Danforth, J. P. Pitts, B. C. Faircloth, P. S. Ward *et al.*, 2017 Phylogenomic Insights into the Evolution of Stinging Wasps and the Origins of Ants and Bees. Current Biology 27: 1019–1025.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST plus : architecture and applications. BMC Bioinformatics 10:.

Cardinal, S., and B. N. Danforth, 2011 The Antiquity and Evolutionary History of Social Behavior in Bees (C. S. Moreau, Ed.). PLoS ONE 6:.

Cha, S. Y., H. J. Yoon, E. M. Lee, M. H. Yoon, J. S. Hwang *et al.*, 2007 The complete nucleotide sequence and gene organization of the mitochondrial genome of the bumblebee, Bombus ignitus (Hymenoptera: Apidae). Gene 392: 206–220.

Chalopin, D., M. Naville, F. Plard, D. Galiana, and J.-N. Volff, 2015 Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. Genome Biol Evol 7: 567–580.

Crozier, R. H., and Y. C. Crozier, 1993 The mitochondrial genome of the honeybee Apis mellifera: complete sequence and genome organization. Genetics 133: 97–117.

Danforth, B. N., 2002 Evolution of sociality in a primitively eusocial lineage of bees. Proc Natl Acad Sci USA 99: 286–290.

Dietz, L., P. Brand, L. M. Eschner, and F. Leese, 2015 The mitochondrial genomes of the caddisflies Sericostoma personatum and Thremma gallicum (Insecta: Trichoptera). Mitochondrial DNA 1–2.

492 Doetterl, S., and N. J. Vereecken, 2010 The chemical ecology and evolution of bee–
493 flower interactions: a review and perspectives. Canadian Journal of Zoology 88:
494 668–697.

495 Elsik, C. G., A. Tayal, C. M. Diesh, D. R. Unni, M. L. Emery *et al.*, 2016 Hymenoptera
496 Genome Database: integrating genome annotations in HymenopteraMine.
497 Nucleic Acids Res. 44: D793–D800.

498 Elsik, C. G., K. C. Worley, A. K. Bennett, M. Beye, F. Camara *et al.*, 2014 Finding the
499 missing honey bee genes: lessons learned from a genome upgrade. BMC
500 Genomics 15: 86.

501 Eltz, T., F. Fritzsch, and J. R. Pech, 2011 Characterization of the orchid bee Euglossa
502 viridissima (Apidae: Euglossini) and a novel cryptic sibling species, by
503 morphological, chemical, and genetic characters. Zool. J. Linn. Soc.

504 Eltz, T., W. M. Whitten, D. W. Roubik, and K. E. Linsenmair, 1999 Fragrance
505 Collection, Storage, and Accumulation by Individual Male Orchid Bees. J. Chem.
506 Ecol. 25: 157–176.

507 Eltz, T., Y. Zimmermann, C. Pfeiffer, J. R. Pech, R. Twele *et al.*, 2008 An olfactory shift
508 is associated with male perfume differentiation and species divergence in orchid
509 bees. Curr. Biol. 18: 1844–1848.

510 Garófalo, C. A., 1985 Social Structure of Euglossa cordata Nests (Hymenoptera:
511 Apidae: Euglossini). Entomologia Generalis 11: 77–83.

512 Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton *et al.*, 2011 High-
513 quality draft assemblies of mammalian genomes from massively parallel
514 sequence data. Proc. Natl. Acad. Sci. U.S.A. 108: 1513–1518.

515 i5K Consortium, 2013 The i5K Initiative: Advancing Arthropod Genomics for
516 Knowledge, Human Health, Agriculture, and the Environment. Journal of
517 Heredity 104: 595–600.

518 Jurka, J., 2000 Repbase Update: a database and an electronic journal of repetitive
519 elements. Trends in Genetics 16: 418–420.

520 Kapheim, K. M., H. Pan, C. Li, S. L. Salzberg, D. Puiu *et al.*, 2015 Genomic signatures of
521 evolutionary transitions from solitary to group living. Science aaa4788.

522 Kidwell, M. G., 2002 Transposable elements and the evolution of genome size in
523 eukaryotes. Genetica 115: 49–63.

524 Klein, A.-M., B. E. Vaissière, J. H. Cane, I. Steffan-Dewenter, S. A. Cunningham *et al.*,
525 2007 Importance of pollinators in changing landscapes for world crops.
526 Proceedings of the Royal Society B: Biological Sciences 274: 303–313.

527 Kocher, S. D., C. Li, W. Yang, H. Tan, S. V. Yi *et al.*, 2013 The draft genome of a socially
528 polymorphic halictid bee, Lasioglossum albipes. Genome Biology 14: R142.

529 Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5:.

530 Laslett, D., and B. Canbäck, 2008 ARWEN: a program to detect tRNA genes in
531 metazoan mitochondrial nucleotide sequences. Bioinformatics 24: 172–175.

532 Leese, F., P. Brand, A. Rozenberg, C. Mayer, S. Agrawal *et al.*, 2012 Exploring
533 Pandora's box: potential and pitfalls of low coverage genome surveys for
534 evolutionary biology. (B. J. Mans, Ed.). PLoS ONE 7: e49202.

535 Litman, J. R., B. N. Danforth, C. D. Eardley, and C. J. Praz, 2011 Why do leafcutter bees
536 cut leaves? New insights into the early evolution of bees. Proc. Biol. Sci. 278:
537 3593–3600.

538 Lowe, T. M., and S. R. Eddy, 1997 tRNAscan-SE: A program for improved detection of
539 transfer RNA genes in genomic sequence. Nucleic Acids Res. 25: 955–964.

540 Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang *et al.*, 2012 SOAPdenovo2: an empirically
541 improved memory-efficient short-read de novo assembler. GigaScience 2016 5:1
542 1: 18.

543 Mayer, C., 2010 Phobos Version 3.3.12. A tandem repeat search program. 20 pp.

544 Menzel, R., and U. Muller, 1996 Learning and Memory in Honeybees: From Behavior
545 to Neural Substrates. Annu. Rev. Neurosci. 19: 379–404.

546 Michener, C. D., 2007 *The Bees of the World. 2nd*. Ed. Johns Hopkins.

547 Nowak, M. A., C. E. Tarnita, and E. O. Wilson, 2010 The evolution of eusociality. 466:
548 1057–1062.

549 Pamilo, P., L. Viljakainen, and A. Vihavainen, 2007 Exceptionally High Density of
550 NUMTs in the Honeybee Genome. Mol Biol Evol 24: 1340–1346.

551 Park, D., J. W. Jung, B.-S. Choi, M. Jayakodi, J. Lee *et al.*, 2015 Uncovering the novel
552 characteristics of Asian honey bee, Apis cerana, by whole genome sequencing.
553 BMC Genomics 16:.

554 Parra, G., K. Bradnam, and I. Korf, 2007 CEGMA: a pipeline to accurately annotate
555 core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067.

556 Pech, M. E. C., W. de J May-Itzá, L. A. M. Medina, and J. J. G. Quezada-Euan, 2008
557 Sociality in Euglossa (Euglossa) viridissima Friese (Hymenoptera, Apidae,
558 Euglossini). 55: 428–433.

559 Pemberton, R. W., and G. S. Wheeler, 2006 Orchid bees don't need orchids: evidence

560 from the naturalization of an orchid bee in Florida. Ecology 87: 1995–2001.

561 Peters, R. S., L. Krogmann, C. Mayer, A. Donath, S. Gunkel *et al.*, 2017 Evolutionary
562 History of the Hymenoptera. Current Biology 0:.

563 Platt, R. N., L. Blanco-Berdugo, and D. A. Ray, 2016 Accurate Transposable Element
564 Annotation Is Vital When Analyzing New Genome Assemblies. Genome Biol Evol
565 8: 403–410.

566 Ramírez, S., R. L. Dressler, and M. Ospina, 2002 Abejas euglosinas (Hymenoptera:
567 Apidae) de la Región Neotropical: Listado de especies con notas sobre su
568 biología. 3: 7.

569 Ramírez, S. R., T. Eltz, M. K. Fujiwara, G. Gerlach, B. Goldman-Huertas *et al.*, 2011
570 Asynchronous Diversification in a Specialized Plant-Pollinator Mutualism.
571 Science 333: 1742–1746.

572 Romiguier, J., S. A. Cameron, S. H. Woodard, B. J. Fischman, L. Keller *et al.*, 2015
573 Phylogenomics controlling for base compositional bias reveals a single origin of
574 eusociality in corbiculate bees. Mol Biol Evol msv258.

575 Roubik, D. W., and P. E. Hanson, 2004 *Orchid bees of tropical America: biology and*
576 *field guide*. Instituto Nacional de Biodiversidad (INBio), Santo Domingo De
577 Heredia.

578 Sadd, B. M., S. M. Barribeau, G. Bloch, D. C. de Graaf, P. Dearden *et al.*, 2015 The
579 genomes of two key bumblebee species with primitive eusocial organization.
580 Genome Biology 16: 76.

581 Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov,
582 2015 BUSCO: assessing genome assembly and annotation completeness with
583 single-copy orthologs. Bioinformatics 31: 3210–3212.

584 Skov, C., and J. Wiley, 2005 Establishment of the neotropical orchid bee Euglossa
585 viridissima (Hymenoptera: Apidae) in Florida. Florida Entomologist 88: 225–
586 227.

587 Slater, G. S. C., and E. Birney, 2005 Automated generation of heuristics for biological
588 sequence comparison. BMC Bioinformatics 6: 31.

589 Soro, A., J. J. G. Quezada-Euan, P. Theodorou, R. F. A. Moritz, and R. J. Paxton, 2016
590 The population genetics of two orchid bees suggests high dispersal, low diploid
591 male production and only an effect of island isolation in lowering genetic
592 diversity. Conserv Genet 1–13.

593 Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler, 2008 Using native and
594 syntenically mapped cDNA alignments to improve de novo gene finding.

595 Bioinformatics 24: 637–644.

596 Stolle, E., L. Wilfert, R. Schmid-Hempel, P. Schmid-Hempel, M. Kube *et al.*, 2011 A
597 second generation genetic map of the bumblebee Bombus terrestris (Linnaeus,
598 1758) reveals slow genome and chromosome evolution in the Apidae. BMC
599 Genomics 12: 48.

600 Suni, S. S., 2016 Dispersal of the orchid bee Euglossa imperialis over degraded
601 habitat and intact forest. Conserv Genet 1–10.

602 Suni, S. S., and B. J. Brosi, 2012 Population genetics of orchid bees in a fragmented
603 tropical landscape. Conserv Genet 13: 323–332.

604 Vogel, S., 1966 Parfümsammelnde Bienen als Bestäuber von Orchidaceen und
605 Gloxinia. Österr bot Z 113: 302–361.

606 Wcislo, W. T., and J. H. Cane, 2003 Floral Resource Utilization by Solitary Bees
607 (Hymenoptera: Apoidea) and Exploitation of Their Stored Foods by Natural
608 Enemies. 41: 257–286.

609 Weinstock, G. M., G. E. Robinson, R. A. Gibbs, K. C. Worley, J. D. Evans *et al.*, 2006
610 Insights into social insects from the genome of the honeybee Apis mellifera.
611 Nature 443: 931–949.

612 Whitten, W. M., A. M. Young, and D. L. Stern, 1993 Nonfloral sources of chemicals
613 that attract male euglossine bees (Apidae: Euglossini). J. Chem. Ecol. 19: 3017–
614 3027.

615 Woodard, S. H., B. J. Fischman, A. Venkat, M. E. Hudson, K. Varala *et al.*, 2011 Genes
616 involved in convergent evolution of eusociality in bees. Proc Natl Acad Sci USA
617 108: 7472–7477.

618 Xu, H., X. Luo, J. Qian, X. Pang, J. Song *et al.*, 2012 FastUniq: a fast de novo duplicates
619 removal tool for paired short reads. (D. Doucet, Ed.). PLoS ONE 7: e52249.

620 Zdobnov, E. M., F. Tegenfeldt, D. Kuznetsov, R. M. Waterhouse, F. A. Simão *et al.*,
621 2017 OrthoDB v9.1: cataloging evolutionary and functional annotations for
622 animal, fungal, plant, archaeal, bacterial and viral orthologs. - PubMed - NCBI.
623 Nucleic Acids Res. 45: D744–D749.

624 Zimmermann, Y., D. L. P. Schorkopf, R. F. A. Moritz, R. W. Pemberton, J. J. G. Quezada-
625 Euan *et al.*, 2011 Population genetic structure of orchid bees (Euglossini) in
626 anthropogenically altered landscapes. Conserv Genet 12: 1183–1194.

627

628

Figures

Figure 1. Genomic features. (A) Phylogeny of the four corbiculate bee tribes with orchid bees as sistergroup to honey bees, stingless bees, and bumblebees (Romiguer et al. 2015). **(B)** K-mer distribution spectrum (k=25) of genomic sequence reads. The positively skewed spectrum reveals a high abundance of a few k-mers, leading to an estimate of 87.7% repetitiveness of the *E. dilemma* genome. Red shows the k-mer spectrum before, and blue after error correction. **(C)** Genomic element density including genic and non-genic features as a fraction of the overall genome assembly length. Over 85% of the assembly could not be annotated with the selected methods. **(D)** Synteny between the *E. dilemma* and the honey bee (*Apis mellifera*) genome. In an analysis including *E. dilemma* scaffolds of ≥100kb length, 83% showed ≥95% synteny to a single honeybee scaffold. Photographs in **(A)** are reproduced from Wikimedia under the CC BY-SA 3.0 license.

Figure 2. Mitochondrial genome reconstruction. The structure of the honey bee mitochondrial genome and information of the homologous reconstructed parts of the *E. dilemma* mitochondrial genome. Non-reconstructed parts of incompletely reconstructed genes are hatched.

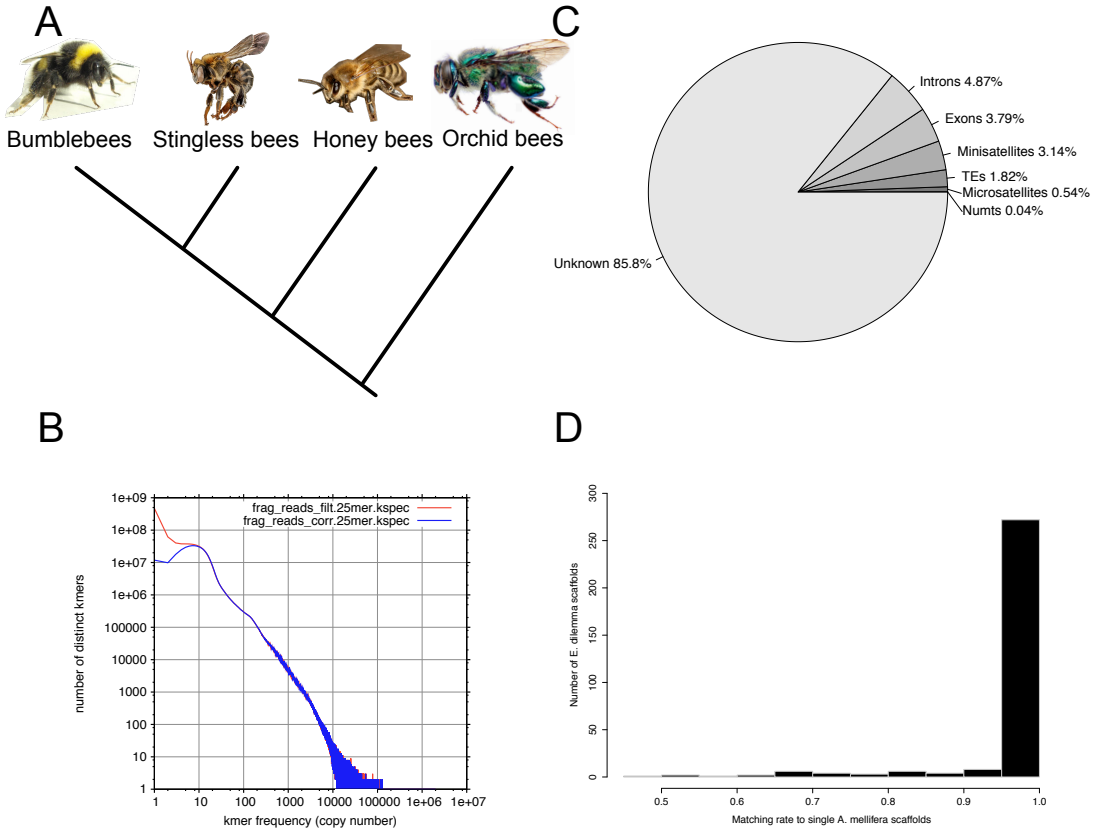

Figure 1

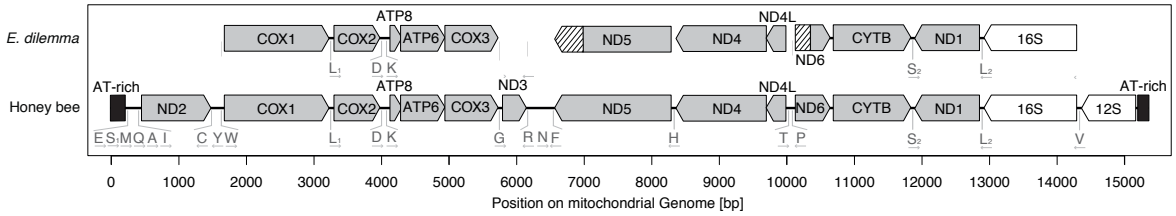

Figure 2

Tables

Table 1. *E. dilemma* genome assembly statistics in comparison to previously published bee genomes.

Table 1

Species	N50	N25	Longest scaffold	Scaffolds	Assembly length	%GC	Predicted genes	Ref.
Euglossa dilemma	143,590	1,417,006	10,108,120	22,698	588,199,720	39.94	15,904	1
Eufriesea mexicana	2,427	443,231	4,677,300	3,522,543	1,031,837,970	41.38	12,022	2
Apis mellifera	997,192	1,922,192	4,736,299	5,644	234,070,657	32.70	15,314	3
Melipona quadrifasciata	68,085	1,896,322	12,087,087	38,604	507,114,161	38.88	14,257	2
Bombus impatiens	1,399,493	2,389,513	5,466,090	5,559	249,185,056	37.75	15,896	4
Lasioglossum albipes	616,426	1,130,413	3,533,895	41,433	341,616,641	41.50	13,448	5

N50 and N25 indicate the length of the shortest scaffold of those including 50% and 25% of the base pairs in a genome assembly. References (Ref.): 1: This study, 2: Kapheim et al. 2015, 3: Elsik et al. 2014, 4: Sadd et al. 2015, 5: Kocher et al. 2012.