Genus-wide characterization of bumblebee genomes reveals variation associated with key ecological and behavioral traits of pollinators

Cheng Sun[1†*], Jiaxing Huang[1†], Yun Wang[2†], Xiaomeng Zhao[1†], Long Su[1†], Gregg W.C. Thomas[3†], Mengya Zhao[4†], Xingtan Zhang[5], Irwin Jungreis[6,7], Manolis Kellis[6,7], Saverio Vicario[8], Igor V. Sharakhov[9,10], Semen M. Bondarenko[9], Martin Hasselmann[11], Chang N Kim[12], Benedict Paten[12], Luca Penso-Dolfin[13], Li Wang[14], Yuxiao Chang[14], Qiang Gao[15], Ling Ma[15], Lina Ma[16], Zhang Zhang[16], Hongbo Zhang[2], Huahao Zhang[17], Livio Ruzzante[18], Hugh M. Robertson[19], Yihui Zhu[20], Yanjie Liu[1], Huipeng Yang[1], Lele Ding[1], Quangui Wang[1], Weilin Xu[1], Cheng Liang[21], Michael W. Itgen[22], Lauren Mee[23], Ben M. Sadd[24], Gang Cao[4], Ze Zhang[2], Matthew Hahn[25], Sarah Schaack[26], Seth M. Barribeau[23], Paul H. Williams[27], Robert M. Waterhouse[18*] and Rachel Lockridge Mueller[22]

†Contributed equally.

***Corresponding authors**: Cheng Sun (suncheng@caas.cn) and Robert M. Waterhouse (robert.waterhouse@unil.ch).

[1]Institute of Apicultural Research, Chinese Academy of Agricultural Sciences, Beijing, China

[2]School of Life Sciences, Chongqing University, Chongqing, China

[3]Division of Biological Sciences, University of Montana, Missoula, Montana, USA

[4]State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan, China

[5]Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, China

[6]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts 02139, USA

[7]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA

[8]Institute of Atmospheric Pollution Research-Italian National Research Council C/O Department of Physics, University of Bari, via Orabona 4, 70125 Bari, Italy

[9]Department of Entomology, Virginia Polytechnic and State University, Blacksburg, VA, 24061, USA

[10]Department of Cytology and Genetics, Tomsk State University, Tomsk 634050, Russian Federation

[11]Department of Livestock Population Genomics, Institute of Animal Science, University of Hohenheim, Stuttgart, Germany

[12]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, USA

[13]German Cancer Research Center, Heidelberg, Germany

[14]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

[15]BGI Genomics, BGI-Shenzhen, Shenzhen, China

[16]China National Center for Bioinformation & Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China
[17]College of Pharmacy and Life Science, Jiujiang University, Jiujiang, China
[18]Department of Ecology and Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[19]Department of Entomology, University of Illinois at Urbana-Champaign, Champaign, IL, USA
[20]Department of Medical Microbiology and Immunology, Genome Center, and MIND Institute, University of California, Davis, CA, USA
[21]Institute of Sericultural and Apiculture, Yunnan Academy of Agricultural Sciences, Mengzi, China
[22]Department of Biology, Colorado State University, Fort collins, CO, USA
[23]Department of Ecology, Evolution and Behaviour, Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom
[24]School of Biological Sciences, Illinois State University, Normal, Illinois, USA
[25]Department of Biology & Department of Computer Science, Indiana University, Bloomington, IN, USA
[26]Department of Biology, Reed College, Portland, Oregon, USA
[27]Department of Life Sciences, Natural History Museum, London, United Kingdom

Abstract

Bumblebees are a diverse group of globally important pollinators in natural ecosystems and for agricultural food production. With both eusocial and solitary life-cycle phases, and some social parasite species, they are especially interesting models to understand social evolution, behavior, and ecology. Reports of many species in decline point to pathogen transmission, habitat loss, pesticide usage, and global climate change, as interconnected causes. These threats to bumblebee diversity make our reliance on a handful of well-studied species for agricultural pollination particularly precarious. To broadly sample bumblebee genomic and phenotypic diversity, we *de novo* sequenced and assembled the genomes of 17 species, representing all 15 subgenera, producing the first genus-wide quantification of genetic and genomic variation potentially underlying key ecological and behavioral traits. The species phylogeny resolves subgenera relationships while incomplete lineage sorting likely drives high levels of gene tree discordance. Five chromosome-level assemblies show a stable 18-chromosome karyotype, with major rearrangements creating 25 chromosomes in social parasites. Differential transposable element activity drives changes in genome sizes, with putative domestications of repetitive sequences influencing gene coding and regulatory potential. Dynamically evolving gene families and signatures of positive selection point to genus-wide variation in processes linked to foraging, diet and metabolism, immunity and detoxification, as well as adaptations for life at high altitudes. These high-quality genomic resources capture natural genetic and phenotypic variation across bumblebees, offering new opportunities to advance our understanding of their remarkable ecological success and to identify and manage current and future threats.

Main Text

Introduction

Bumblebees (Hymenoptera: *Apidae*) are a group of pollinating insects comprising the genus *Bombus*, which are economically important for crop pollination (Garibaldi, et al. 2013; Martin, et al. 2019; Velthuis and Van Doorn 2006). Bumblebees are also ecologically important pollinators, serving as the sole or predominant pollinators of many wild plants (Fontaine, et al. 2005; Goulson, et al. 2008). They are particularly charismatic social insects that exhibit complex behaviors such as learning through observation (Alem, et al. 2016) and damaging leaves to stimulate earlier flowering (Pashalidou, et al. 2020). Global and local environmental changes have resulted in some species declining in abundance and others remaining stable or even increasing (Bartomeus, et al. 2013; Cameron, et al. 2011; Cameron and Sadd 2019; Koch, et al. 2015). Decline in bumblebee abundance and distribution resulting from habitat loss, pathogen transmission, climate change, and agrochemical exposure is threatening pollination services to both wild plants and crops, raising concerns for bumblebees, the plant species they service, food security, and ecosystem stability (Cameron and Sadd 2019; Goulson, et al. 2015; Grixti, et al. 2009; Potts, et al. 2010; Soroye, et al. 2020; Williams and Osborne 2009).

Bumblebees comprise ~250 extant species classified into 15 subgenera (Williams 1998; Williams, et al. 2018). The initial diversification of *Bombus* lineages occurred ~25–40 million years ago (Ma), near the Eocene-Oligocene boundary ~34 Ma (Hines 2008; Williams 1998). Bumblebees display considerable interspecific diversity in morphology, food preference, pathogen incidence, and exhibit diverse life histories and ecologies (Arbetman, et al. 2017; Persson, et al. 2015; Sikora and Kelm 2012; Williams 1994). Members of the subgenus *Mendacibombus*, the sister group to all other extant bumblebees, are high-elevation specialists with distributions centered on the Qinghai-Tibetan plateau (Williams, et al. 2018). Species in the subgenus *Psithyrus* exhibit social parasitism; they do not have a worker caste, and they feed on food collected by host workers (Lhomme and Hines 2019). Bumblebees are distributed across the globe, from Greenland to the Amazon Basin and from sea level to altitudes of 5,640 m in the Himalayas, where they occupy diverse habitats, from alpine meadows to lowland tropical forest (Williams and Paul 1985; Williams, et al. 2018). Much remains to be learned about bumblebees. For example, little is known about the underlying genetic and genomic variation that gives rise to these diverse phenotypes, including their differential responses to changing environments.

To broadly sample this genomic and phenotypic diversity, we performed *de novo* sequencing and assembly of the genomes of 17 bumblebee species, representing all of the 15 subgenera within the genus *Bombus*. Integrating these datasets with two previously published bumblebee genomes, we performed comparative analyses of genome structures, genome contents, and gene evolutionary dynamics across the phylogeny. Our results characterizing bumblebee gene and genome evolution provide the first genus-wide quantification of genetic and genomic variation potentially underlying key eco-ethological traits.

Results

High quality genomic resources for all 15 *Bombus* subgenera

Sequencing and assembly strategies resulted in high quality genomic resources with 12 scaffold-level and five chromosome-level genome assemblies (Table 1). Criteria including phylogenetic position, species traits, and geographic distribution were applied to select species for whole genome sequencing from across the genus. For the five species for which sufficient samples could be collected, high-throughput chromatin conformation capture (Hi-C) (Belton, et al. 2012) was used to produce chromosome-level genome assemblies (Table 1). A total of 17 species were selected (Additional file 1: Table S1), which span all 15 subgenera within *Bombus* (Williams, et al. 2008). Among these, two species (*B. superbus* and *B. waltoni*) are from *Mendacibombus*, the earliest split in the *Bombus* phylogeny; four species (*B. superbus*, *B. waltoni*, *B. skorikovi*, and *B. difficillimus*) inhabit high elevations (> 4000 m above sea level); two species (*B. turneri* and *B. skorikovi*) exhibit social parasitism; and one species (*B. polaris*) is endemic to Arctic/subarctic regions (Williams, et al. 2019). In addition, species traits including range size, tongue length, parasite incidence, and decline status vary across the selected species (Arbetman, et al. 2017; Williams 1994)(Additional file 1: Table S1).

Table 1. Genome assembly results of the 17 newly sequenced species.

	Contig Size (Mb)	Contig N50 (Kb)	Scaffold Size (Mb)	Scaffold N50 (Mb)	Chromosome Size (Mb)	Chromosome N50 (Mb)
B. superbus	229.84	441.61	230.16	6.90	NA	NA
B. waltoni	230.89	430.54	231.17	4.66	NA	NA
B. confusus	238.52	227.26	239.12	3.26	NA	NA
B. haemorrhoidalis	239.34	572.47	239.59	4.74	240.54	15.09
B. ignitus	240.60	374.12	241.36	3.02	242.57	15.19
B. skorikovi	241.25	225.53	242.05	4.34	NA	NA
B. opulentus	241.99	267.78	242.38	2.42	NA	NA
B. turneri	242.39	212.53	243.01	4.34	243.11	9.70
B. soroeensis	243.19	244.99	243.68	2.12	NA	NA
B. polaris	245.17	152.35	245.82	2.25	NA	NA
B. breviceps	246.03	578.55	246.41	4.04	248.12	14.71
B. cullumanus	246.56	422.80	247.01	4.58	NA	NA
B. difficillimus	247.45	177.31	248.33	2.07	NA	NA
B. consobrinus	248.56	284.90	249.09	4.77	NA	NA
B. pyrosoma	251.86	472.32	252.70	6.07	254.80	15.22
B. picipes	253.31	185.91	254.01	5.88	NA	NA
B. sibiricus	261.72	253.94	262.49	3.14	NA	NA

Note: Mb, megabase; Kb, kilobase; NA, not applicable.

Sequencing and assembly strategies included generating two Illumina sequencing datasets for each species: (i) overlapping paired-end reads (2 × 250 bp) from one small-insert fragment library (insert size: 400 or 450 bp); and (ii) paired-end reads (2 × 150 bp) from four large-insert jump libraries (insert sizes: 4 kb, 6kb, 8kb and 10 kb, respectively; Additional file 1: Table S2). Whole-genome overlapping paired-end reads

from fragment libraries were assembled into continuous sequences (contigs) using the software DISCOVAR *de novo* (Love, et al. 2016), then scaffolded with reads from jump libraries using the software BESST (Sahlin, et al. 2014). The resulting assemblies have a mean contig N50 of 325 Kb, ranging up to 579 Kb for *B. breviceps*; the mean scaffold N50 is 4.0 Mb, ranging up to 6.9 Mb for *B. superbus* (Table 1). Genome assembly quality in terms of expected gene content was evaluated by Benchmarking Universal Single-Copy Ortholog (BUSCO) analysis (Waterhouse, et al. 2018), which showed high BUSCO completeness scores (average 99.0%, from 97.5 to 99.6%; Additional file 2: Figure S1) for all genomes.

Genome annotation resulted in total protein-coding gene predictions per species ranging from 14,027–16,970 (mean = 15,838, standard deviation = 908; Additional file 1: Table S3). These were annotated using the MAKER pipeline (Cantarel, et al. 2008), based on *ab initio* gene predictions, transcript evidence, and homologous protein evidence. Gene counts are similar to those of 12 drosophilid species (mean = 15,361, sd. = 852 Clark et al., 2007), but are higher than those of 19 anophelines (mean = 13,110, sd. = 1,397) (Neafsey, et al. 2015), and they do not correlate with assembly contiguity ($p = 0.1757$; Additional file 2: Figure S2). Between 7,299–8,135 genes were assigned at least one Gene Ontology (GO) term and 9,431–10,578 genes were annotated with at least one protein domain (Additional file 1: Table S3). BUSCO analysis of the annotated genes also showed high completeness scores for all species (Additional file 2: Figure S3). Furthermore, comprehensive miRNA, tRNA, and lncRNA gene prediction revealed an average of 93, 306, and 3,353 genes, respectively (Additional file 1: Table S3). Finally, transposable element (TE) annotation showed that the total TE content ranged from 9.66% (22.2 Mb) in *B. superbus* to 17.88% (46.9 Mb) in *B. sibiricus* (Additional file 1: Table S4).

Genome-scale phylogeny of bumblebees

The species-level molecular phylogeny (Figure 1A) estimated from maximum-likelihood analysis with IQTree (Minh, et al. 2020b) is largely consistent with previously inferred phylogenetic relationships of the 15 subgenera based on five genes (Cameron, et al. 2007; Williams, et al. 2008), showing only two topological differences. The results support previous conclusions that: (i) subgenus *Mendacibobus* (*Md*) is the sister group to all the other subgenera; and (ii) lineages named *Psithyrus (Ps)* are within the *Bombus* clade, arguing they should not be named as an independent genus (Figure 1A). The species phylogeny was built from the concatenated aligned protein sequences of 3,617 universal single-copy orthologs from 19 bumblebee species (17 from the current study, two published previously: *B. terrestris* and *B. impatiens* (Sadd, et al. 2015)) and four honeybee species (*A. florea*, *A. dorsata* (Oppenheim, et al. 2020), *A. cerana* (Park, et al. 2015), and *A. mellifera* (Weinstock, et al. 2006)), with orthologous groups delineated using the OrthoDB software (Kriventseva, et al. 2015). Complementary analysis with ASTRAL (Zhang, et al. 2018) resulted in an identical species tree with the exception of the placement of *B. pyrosoma*, which no longer forms a monophyletic pairing with *B. breviceps* (Additional file 2: Figure S4). This type of discordance between species tree methods is consistent with a known shortcoming of maximum-likelihood concatenation in the presence of incomplete lineage sorting (ILS) (Kubatko and Degnan 2007; Mendes and Hahn 2018), implying that the ASTRAL topology is likely the correct topology.

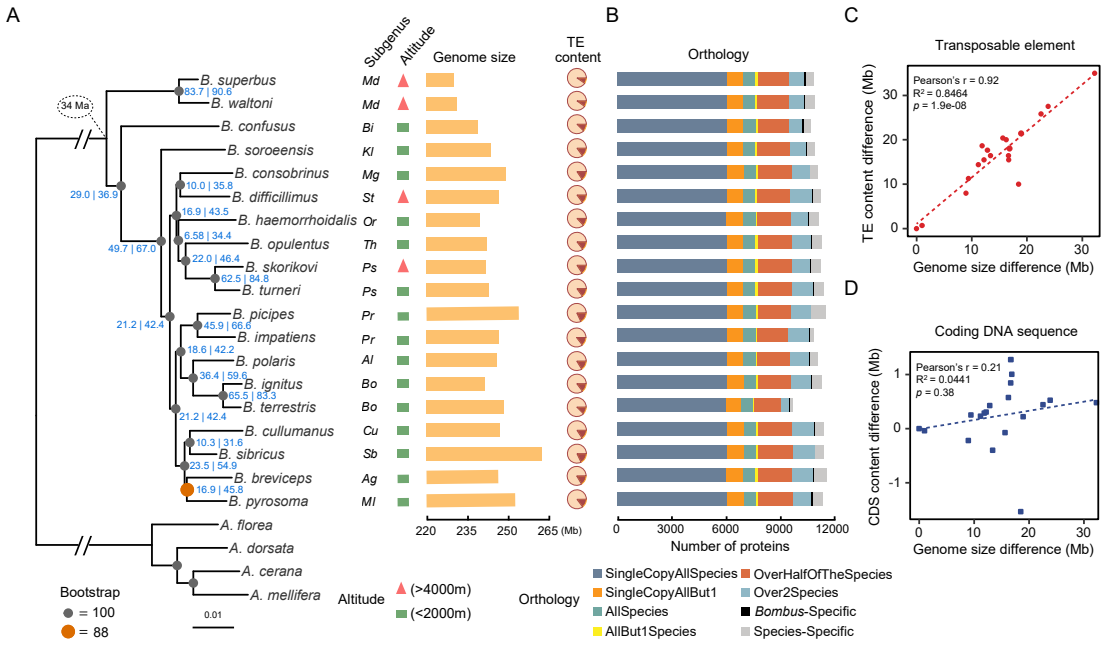

Figure 1. Phylogenetic, genomic and proteomic comparisons of 19 bumblebee species. (**A**) From left to right: the maximum likelihood molecular species phylogeny built from 3,617 concatenated single-copy orthologous groups from all sequenced bumblebees and honeybee outgroups. Node labels in blue are of the following format: gene concordance factors | site concordance factors. Branches scaled by relative number of substitutions; the subgenus that each bumblebee species belongs to (Md, *Mendacibombus*; Bi, *Bombias*; Kl, *Kallobombus*; Mg, *Megabombus*; St, *Subterraneobombus*; Or, *Orientalibombus*; Th, *Thoracobombus*; Ps, *Psithyrus*; Cu, *Cullumanobombus*; Sb, *Sibiricobombus*; Ag, *Alpigenobombus*; Ml, *Melanobombus*; Pr, *Pyrobombus*; Al, *Alpinobombus*; Bo, *Bombus*); altitude of species collection site (red triangle: extreme high-altitude; green rectangle: low-altitude); genome assembly size of each sequenced species; fraction of transposable elements (TE) (brown) in each genome. (**B**) Bar plots show total gene counts for each bumblebee partitioned according to their orthology profiles, from ancient genes found across bumblebees to lineage-restricted and species-specific genes. (**C**) and (**D**) represent the contribution of transposable element and coding DNA sequence to genome size variation across bumblebees, respectively. Differences in the total content of transposable elements (**C**) and coding DNA sequences (**D**) of the 19 genomes relative to that of *B. superbus* (which has the smallest genome assembly size) are plotted against their genome size differences (relative to that of *B. superbus*).

However, inferring rooted gene trees from 3,530 single-copy orthologous groups reveals extreme levels of discordance: none of their topologies match the topology of the tree inferred from concatenation (Additional file 1: Table S5 and Additional file 1: Table S6), and nearly every gene tree has a unique topology (Additional file 1: Table S7). Such extreme levels of discordance have been observed previously in birds (Jarvis, et al. 2014) and tomatoes (Pease, et al. 2016), and have been attributed to a variety of sources, such as ILS and introgression (Maddison 1997). A lack of informative sites, only 24%, compared to 47% in a similar dataset of 25 drosophilids (Da Lage, et al. 2019), possibly due to the relatively recent diversification of bumblebees (Hines 2008), may also cause discordance. Concordance analysis (Minh, et al. 2018) shows that, on average, nodes in the species tree are present in only a third of gene trees and only about half of informative sites support the species tree (node labels in Figure 1A). These site concordance factors, the short internal branches of the species tree, and the strong correlation between them (Additional file 2: Figure S5), are consistent with ILS driving the observed gene tree discordance. Gene-level phylogenies are therefore used in all subsequent gene-based molecular evolution analyses because such discordance can bias inferences of substitutions when mapped onto a species tree (Mendes and Hahn 2016).

Major genomic rearrangements in social parasites

The five Hi-C genome assemblies indicate that four of the five subgenera have 18 chromosomes (Figure 2A and 2C; Additional file 2: Figure S6A-B), consistent with previous karyotypic analysis that inferred the ancestral chromosome number is 18 (Owen, et al. 1995). However, the social parasite bumblebee, *B. turneri,* subgenus *Psithyrus*, has 25 chromosomes (Figure 2B), consistent with previous cytological work (Owen and Robin 1983). Despite the higher chromosome number, its genome size is within the range of other bumblebees (Figure 1; Table 1). Pairwise comparisons between *B. turneri* and each of the other four chromosomal-level assemblies to investigate macrosynteny relationships and understand how a 25-chromosome karyotype was derived from the ancestral state revealed three processes. First, some chromosomes descended, structurally unchanged, from ancestral chromosomes (e.g., chromosome 5; Figure 2D in blue). Second, some originated by fission of an ancestral chromosome (e.g., 11 and 25 of *B. turneri* originated by the fission of ancestral chromosome 11; Figure 2D in red). Lastly, some are derived from fusions of two or more ancestral chromosome segments (e.g., *B. turneri* chromosome 22 was derived from the fusion of segments of ancestral chromosomes 7, 8, 10, and 16 (Figure 2D in gold). Pairwise comparisons between *Psithyrus* and members of other subgenera reveal similar results, and support the inference that the 25 chromosomes of the social parasite bumblebee result from a combination of fission, fusion, and retention of ancestral chromosomes (Additional file 2: Figure S6).

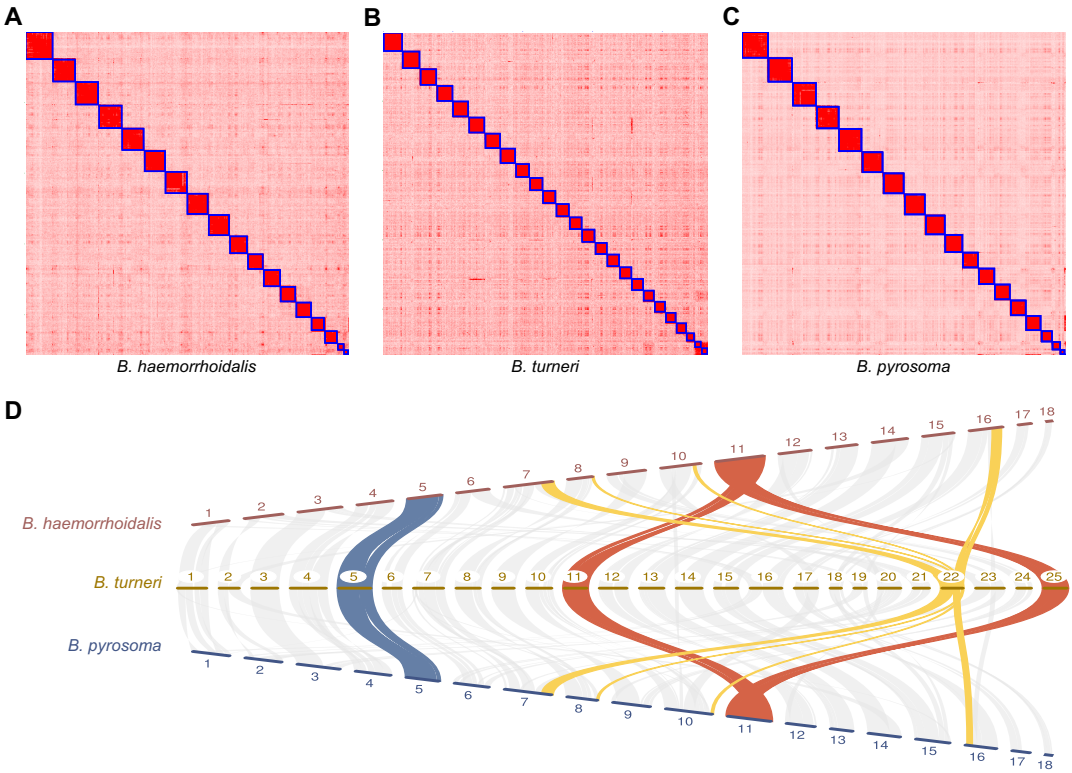

Figure 2. Chromosome number evolution in representative *Bombus* species. Hi-C contact heatmaps for *B. haemorrhoidalis* (**A**), *B. turneri* (**B**), and *B. pyrosoma* (**C**) show that the three species have 18, 25, and 18 chromosomes, respectively. The 18-chromosome karyotype is the inferred ancestral genome structure, with 25 chromosomes found in social parasite bumblebees of the subgenus *Psithyrus*. (**D**) Macrosynteny comparisons across *B. haemorrhoidalis*, *B. turneri* and *B. pyrosoma* shows how the 25 *B. turneri* chromosomes result from a combination of fission (red), fusion (yellow), and retention (blue) of ancestral chromosomes.

Rates of chromosome evolution, in terms of rearrangements relative to *B. terrestris,* were investigated for each of the five species with chromosome-level assemblies. Rearrangement rates in bumblebees range from 0.0016–0.0075 inversions/Mb/My (Additional file 1: Table S8), which is much lower than those of drosophilids (0.013–0.159 inversions/Mb/My) and anophelines (0.052–0.068 inversions/Mb/My) (Neafsey, et al. 2015; von Grotthuss, et al. 2010). Thus, although bumblebee genomes have a high recombination rate (Wilfert, et al. 2007), their rates of chromosome evolution are relatively slow, which is further supported by the observed high synteny contiguity across species (average 88%, from 80–95%; Additional file 1: Table S9).

Transposable elements drive genome size variation

Genome assembly sizes (haploid) range from 230 Mb in *B. superbus* to 262 Mb in *B. sibiricus* (Figure 1). Ancestral genome size inference of bumblebees produced an estimate of 230-231 Mb, similar to that of members of the subgenus *Mendacibombus*, but smaller than the genomes of all other extant bumblebees surveyed (Additional file 2: Figure S7).

Comparing genome size differences with relative content of TEs, simple sequence repeats (SSRs), and coding DNA sequences (CDS) shows that TE content explains a majority of the differences across bumblebees (Pearson correlation R = 0.92, *P* = 1.9e-08, R^2=0.85; Figure 1C, Figure 1D, Additional file 2: Figure S8). *Mendacibombus* species have a smaller genome size than other species (Figure 1), and TEs that transposed in non-*Mendacibombus* species after divergence from *Mendacibombus* show copy numbers ranging from 1,992–4,755 (Additional file 2: Figure S9), supporting the contribution of TEs to genome size evolution. Furthermore, TE proliferation history analysis indicated that all non-*Mendacibombus* species have more recent TE amplification peaks (Additional file 2: Figure S10), consistent with increased TE activity driving genome size increases.

The genomic distributions of TEs include 1,074–1,786 TE loci that overlap with the coding regions of protein-coding genes (Additional file 1: Table S10), and 352 of these genes are universal single-copy across the 19 bumblebees whose *dN/dS* values are all < 1 (Additional file 1: Table S11), indicating that TEs may have been exonized in bumblebee genomes to form novel proteins. In addition, there are thousands of TEs located within 1 kb of a gene in each species (Additional file 1: Table S10), and, in *B. terrestris*, 278 such TEs co-locate with open chromatin regions detected by ATAC-seq (Additional file 1: Table S12), suggesting those TEs may have become incorporated into regulatory sequences.

Gene content evolution reflects foraging and diet diversity

Orthology delineation results indicate that a majority of genes are found in one or more copies in nearly all lineages across bumblebees (Figure 1B). These include 53 *Bombus*-specific ortholog groups, which are present in all 19 bumblebees but absent in all four honeybees (Figure 1B; Additional file 1: Table S13), and may play roles in lineage-specific traits. Functional annotation suggests that five of these *Bombus*-specific genes are associated with protein metabolism and transport (Additional file 1: Table S13), potentially linked to the higher protein content of pollen collected by bumblebees than honeybees (Leonhardt and Blüthgen 2011). Ortholog groups with the broadest species representation are functionally enriched for core biological processes such as protein transport, signal transduction (e.g. Wnt pathway),

(de)ubiquitination, and cytoskeleton organization (Additional file 1: Table S14). In contrast, those with sparse or lineage-restricted species representation are enriched for processes including smell and taste perception, amino acid biosynthesis, and oxidation-reduction (Additional file 1: Table S14). On average, 465 species-specific genes (those without an ortholog in any other lineage) were identified in each bumblebee species (range 137–767) (Additional file 1: Table S15), which may contribute to species-specific traits but whose functional roles remain to be explored.

Turnover analysis of gene repertoires across the *Bombus* phylogeny (15 species, one per subgenus) using CAFE v3.0 (Han, et al. 2013) identified expansions and contractions among 13,828 gene families and quantified variations in gene gain/loss rates across species (Additional file 2: Figure S11). After error correction, the overall rate of gene gain/loss in *Bombus* genomes is 0.0036/gene/million years, similar to an analysis of 18 anopheline species and 25 drosophilids (Additional file 1: Table S16) (Da Lage, et al. 2019; Neafsey, et al. 2015). However, these genus-specific gene gain/loss rates are 2-3 times higher than order-wide rates, which average 0.0011 (Additional file 1: Table S16) (Thomas, et al. 2020), possibly due to the denser sampling in genus-level studies that allow more events to be captured. Gene gain and loss events, along with the number of rapidly evolving gene families, are summarized for each species (Additional file 1: Table S17), with a total of 3,797 rapidly changing gene families. The most dynamic gene families are enriched for processes including smell and taste perception, chitin metabolism, microtubule-based movement, and methylation (Additional file 1: Table S18). Complementary analysis using three measures of gene copy number variation also identifies these processes as enriched among the most variable gene families, in contrast to the most stable that are involved in processes related to translation, adhesion, and transport (Additional file 1: Table S19). In terms of protein domain copy number evolution, the most highly variable genes are those with protein-protein interaction mediating F-box domains, putatively DNA-binding SAP motifs, and phosphate-transferring guanylate kinases (Additional file 1: Table S20).

Stable intron-exon structures with abundant stop-codon readthrough

Protein-coding potential analysis using *B. terrestris* as the reference species identified 851 candidate readthrough stop codons (Additional file 2: Figure S12; Additional file 1: Table S21), i.e. where translation likely continues through stop codons to produce extended protein isoforms. Coding potential was assessed using PhyloCSF (Lin, et al. 2011) on whole genome alignments of all 19 bumblebees and four honeybees. The false discovery rate was estimated using enrichment for the TGA-C stop codon context, which is favored in readthrough genes, to infer that no more than 30% of the 200 highest-scoring candidates are false positives, and that at least 306 of our 851 candidates undergo functional readthrough. While rare beyond Pancrustacea, hundreds of *Drosophila* and *Anopheles* genes undergo readthrough, and in Hymenoptera estimates for honeybee are low but for *Nasonia* wasps high (Dunn, et al. 2013; Jungreis, et al. 2016; Jungreis, et al. 2011; Rajput, et al. 2019). These whole-genome-alignment-based results support the prediction (Jungreis, et al. 2011) that insect species have abundant stop-codon readthrough.

In contrast, intron-exon boundaries within *Bombus* genes are relatively stable. Examining evolutionary histories of intron gains and losses revealed few changes, representing only 3-4% of ancestral intron sites, with more gains than losses

(Additional file 2: Figure S13; Additional file 1: Table S22), unlike drosophilids and anophelines where losses dominate (Neafsey, et al. 2015), suggesting that bumblebee gene structure has remained relatively stable over the 34 million years since their last common ancestor.

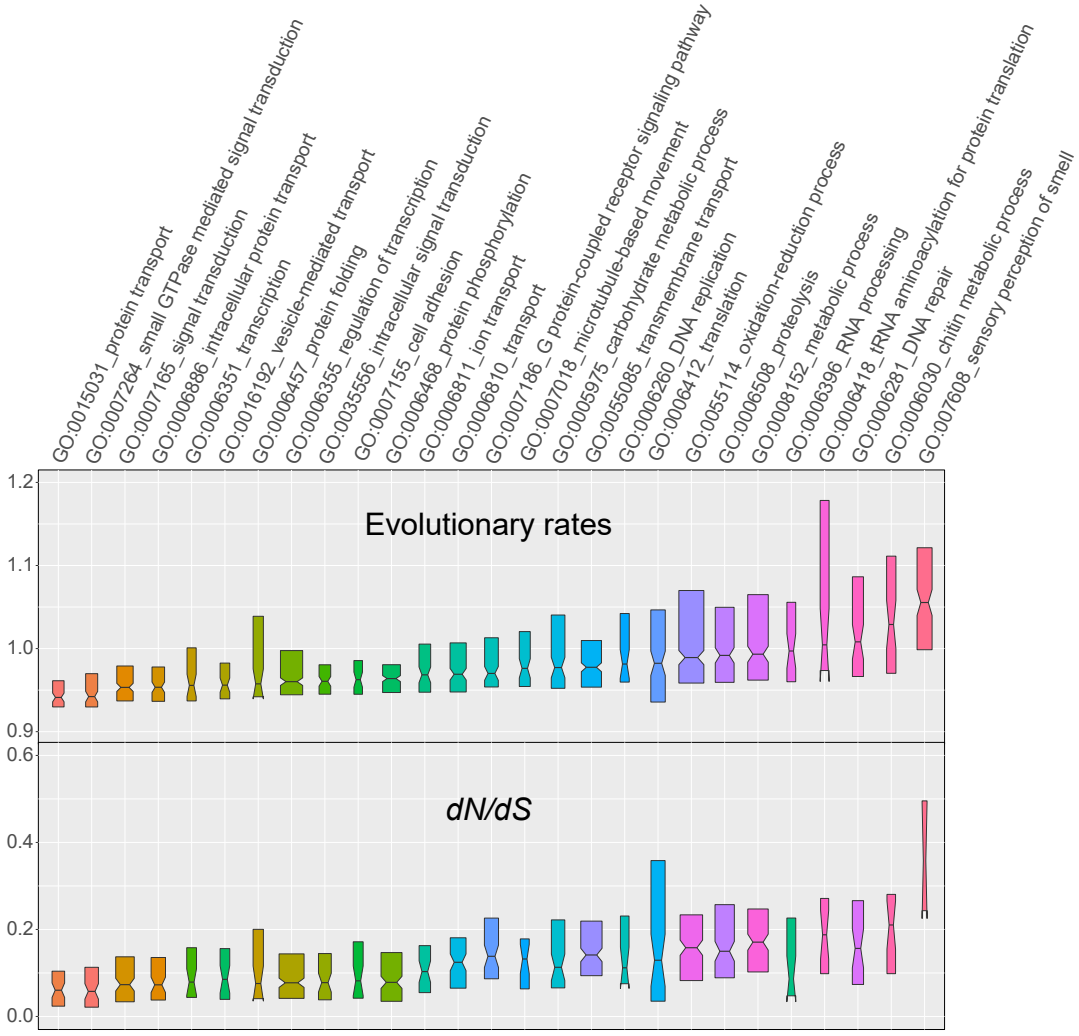

Figure 3. Molecular evolution of protein-coding genes in terms of evolutionary rate (amino acid sequence divergence) and *dN/dS* ratio among selected Gene Ontology (GO) Biological Process terms. Categories are sorted by evolutionary rate from the most conservative (left) to the most dynamic (right) and colored from the highest values (red) to the median value (blue) to the lowest values (orange). Notched boxes show medians of orthologous group values with the limits of the upper and lower quartiles, and box widths are proportional to the number of orthologous groups in each category.

Divergence and selective constraints of protein-coding genes

Bumblebee genes with elevated sequence divergence and/or relaxed constraints include processes related to smell perception, chitin metabolism, RNA processing, DNA repair, and oxidation-reduction (Figure 3). Measures of evolutionary rate (amino acid sequence divergence) and selective constraint (*dN/dS*) showed similar trends among different functional categories of genes. Most genes are strongly constrained, with median estimates of *dN/dS* much lower than one. Assignment of GO terms and InterPro domains is usually biased towards slower-evolving, well-conserved genes (Additional file 2: Figure S14). Nevertheless, functional categories with the fastest-evolving genes are further supported and complemented by examining molecular function GO terms (Additional file 2: Figure S15A) and InterPro domains (Additional file 2: Figure S15B), which show elevated rates for odorant binding, olfactory receptor activity, chitin binding, oxidoreductase activity,

serine-type endopeptidase activity, and olfactory receptor domains. GO term enrichment analysis of the slowest and fastest evolving subsets of genes, bottom and top 20% respectively (Additional file 2: Figure S16), showed genes with the slowest evolutionary rates and the lowest *dN/dS* ratios were enriched for essential house-keeping biological processes and molecular functions (Additional file 1: Table S23; Additional file 1: Table S24). In contrast, genes with the fastest evolutionary rates were enriched for processes linked to polysaccharide biosynthesis, tRNA aminoacylation, drug binding and RNA methyltransferase activity (Additional file 1: Table S23). Genes with the highest *dN/dS* ratios were enriched for processes and functions including proteolysis, translation, ncRNA processing, and chitin metabolism (Additional file 1: Table S24).

Codon usage bias driven by AT content

Analysis of codon usage bias showed no evidence for selection on optimal codons, in contrast to drosophilids but similar to anophelines (Neafsey, et al. 2015; Vicario, et al. 2007). Instead, codon usage bias in bumblebees seems to be driven mainly by AT content, consistent with previous reports in Hymenoptera (Behura and Severson 2012). Optimal codons were estimated in each species and correlation coefficients were computed between relative synonymous codon usage (RSCU) and effective number of codons (ENC) per gene. All species have a similar preference and intensity of preference; for each amino acid, there was a consistently highly preferred codon and often a secondarily preferred one, all ending in A/T (Additional file 2: Figure S17). To test if codon usage could largely be explained by mutation bias, a linear model was used to predict Fop (frequency of optimal codon) from overall gene AT content and amino acid use. The model explained 99.2% of the Fop variation without the need to include the species origin of each gene. The AT content alone explained 81% of the variation (Additional file 2: Figure S18). Moreover, a strong correlation was observed between codon AT content and the correlation between RSCU and ENC across all species (Additional file 2: Figure S19).

Evolution of genes associated with bumblebee eco-ethology

Many ecological and environmental factors—for example, shortage of food, pathogen emergence, pesticide exposure, and climate change—are contributing to the overall decline of bumblebees worldwide (Cameron and Sadd 2019; Goulson, et al. 2015; Williams, et al. 2009). To begin to explore the complement of genes likely to be involved in bumblebee interactions with their environment, we examined the evolution of gene families associated with their ecology and life histories. Sampling across the *Bombus* genus enabled the first survey of natural gene repertoire diversity of such families that are likely to be important for bumblebee adaptability and success.

Chemosensory receptor diversity: Chemosensation plays a critical role in locating food and nests, communicating with nestmates, and identifying other environmental cues (Ayasse and Jarau 2014). A search of the three major chemosensory receptor gene families—odorant receptors (ORs), gustatory receptors (GRs), and ionotropic receptors (IRs)—in the sequenced bumblebee genomes identified 3,228 genes (Additional file 1: Table S25). Only complete genes were used for gene gain and loss analysis. Despite the similarities in total OR gene counts, examples of gene gain/loss were observed in specific lineages. There was a net loss of 15 ORs in the common ancestor of the subgenus *Mendacibombus* (*Md*) (Figure 4A; Additional file 2: Figure S20). Species in *Mendacibombus* mainly inhabit high mountains including the

Qinghai-Tibetan plateau, with relatively low floral diversity (Williams, et al. 2018), which may be linked to OR loss in this lineage. A net loss of 11 ORs was observed in the common ancestor of subgenus *Psithyrus* (*Ps*) (Figure 4A; Additional file 2: Figure S20). For ORs shared across bumblebees, eight showed evidence of positive selection in a subset of species, including putative pheromone receptors (Additional file 1: Table S26). Compared with ORs, GRs and IRs have much lower and more stable gene counts (Additional file 2: Figure S20). However, despite overall conservation of gene number and widespread evidence for purifying selection, there is evidence that some GR and IR genes experienced positive selection in a subset of species, including receptors putatively involved in sensing fructose and temperature (Additional file 1: Table S26).

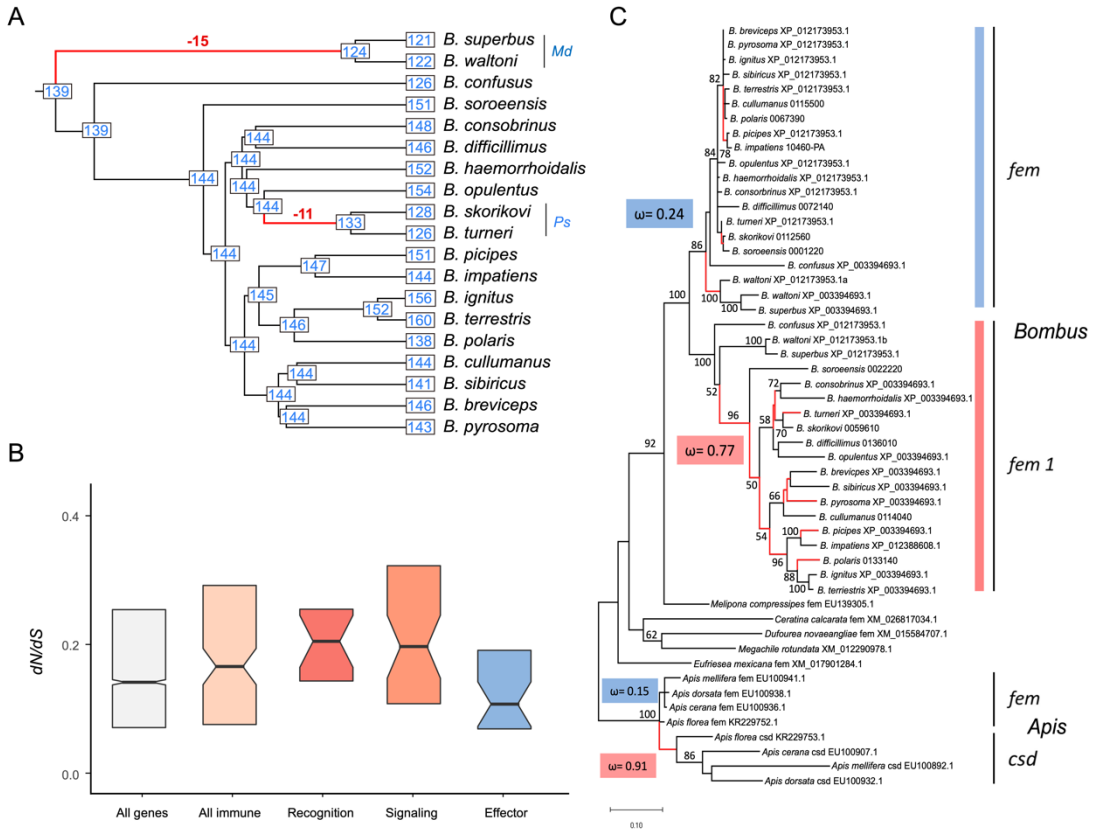

Figure 4. Evolution of genes associated with ecology and reproduction. (**A**). Observed gene counts and inferred ancestral gene counts of bumblebee odorant receptors (ORs) on an ultrametric phylogeny, highlighting two major gene loss events (the complete result is available in Additional file 2: Figure S21). *Md*, *Mendacibombus*; *Ps*, *Psithyrus*. (**B**). Boxplots showing *dN/dS* ratios for different categories of immune genes and all single-copy genes in bumblebee (All genes). Elevated *dN/dS* ratios among immune-related genes is driven by higher ratios for genes involved in recognition and signaling processes. Notched boxes show medians of orthologous group values with the limits of the upper and lower quartiles. (C). The evolutionary history of fem genes of bees including their paralogs *fem1* in *Bombus* and *csd* in *Apis*. Global non-synonymous to synonymous rate ratio (ω) were calculated for *femBombus* (reference, blue) and *fem1Bombus* (test, red), including a branch-site testing framework with model fitting and Likelihood Ratio Tests, showing evidence for relaxation of selection in *fem1Bombus* ($P<0.001$, LR = 36.34). Spurious actions of diversifying selection on branches predominantly found in *fem1Bombus* are marked in red. For comparison, ω for fem and *csd* in *Apis* is given, known as striking example of neo-functionalization.

Detoxification capacity: Detoxification genes are used to neutralize toxic plant secondary metabolites and pesticides. Repertoires of carboxyl/cholinesterases (CCEs), cytochrome P450 monooxygenases (P450s), and glutathione S-transferases (GSTs) in the 17 genomes are much smaller than in drosophilids and anophelines (Additional

file 1: Table S27), indicating a genus-wide deficit of this gene category, previously observed in two bumblebees (Sadd, et al. 2015). There are 88 detoxification genes on average in bumblebees, with little variation across species (Additional file 1: Table S27). Despite overall conservation of gene number and widespread evidence for purifying selection (mean *dN/dS* is 0.26), a total of 30 detoxification genes, including CCEs, P450s, and GSTs, showed evidence of positive diversifying selection in a subset of species (Additional file 1: Table S28).

Immune defense: Immune genes are involved in recognition of and defense against pathogens. Similar to detoxification genes, counts in the 17 sequenced genomes are much lower than in drosophilids and anophelines (Additional file 1: Table S29), showing that the previously noted paucity in two bumblebees (Barribeau, et al. 2015; Sadd, et al. 2015) extends to the whole genus. Bumblebee genomes contain components of all major immune pathways described in insects, and gene counts are fairly conserved across species (Additional file 1: Table S29). For example, all species have two genes encoding Gram-negative bacteria binding-proteins, while peptidoglycan-recognition proteins are more variable with between four and six gene copies. Comparing *dN/dS* ratios between immune genes and all single-copy orthologous genes in bumblebees showed that immune genes exhibit slightly higher *dN/dS* ratios ($P = 0.04$, Wilcoxon rank sum test), and among immune genes, recognition and signaling genes have higher *dN/dS* ratios than effector genes (Figure 4B). In addition, despite widespread evidence for purifying selection, a total of 77 immune genes showed evidence of positive selection in a subset of bumblebee species (Additional file 1: Table S30). *B. cullumanus*, *B. difficillimus*, and *B. confusus*, which have no reported internal parasites (Arbetman, et al. 2017), are among the species that have the most positively selected genes (Additional file 1: Table S30).

Genes involved in high-elevation adaptation

Bombus superbus, *B. waltoni*, *B. difficillimus*, and *B. skorikovi* are four species collected at elevations > 4,000 m that represent three subgenera (Figure 1). No genes show signatures of positive selection in all high-elevation species but none of the low-elevation species. However, nine genes show evidence of positive selection in species representing two of the three high-elevation subgenera, but none of the low-elevation species (Additional file 1: Table S31). Two encode Myosin-VIIa and CPAMD8, respectively, which are involved in eye development (Cheong, et al. 2016; Williams and Lopes 2011). As bumblebees detect flowers visually (Meyer-Rochow 2019), signatures of selection might be related to fine tuning eye development for optimal foraging in high altitude light conditions. Three genes encode histone deacetylase, synaptotagmin-12, and heterogeneous nuclear ribonucleoprotein, which are involved in maintaining muscle integrity and keeping “flight state”, which is critical for undertaking long-distance food-searching (Liu, et al. 2001; Manjila, et al. 2019; Pigna, et al. 2019). Three genes encode sodium-coupled monocarboxylate transporter, glycosyltransferase family protein, and xyloside xylosyltransferase 1, these genes are believed to be involved in metabolic adaptation to hypoxia (Gustafsson, et al. 2005; Lee, et al. 2013; Shirato, et al. 2010; Véga, et al. 2006) (Additional file 1: Table S31). The remaining gene encodes a proton channel, which may be also involved in the metabolic adaptation to hypoxia (Bacon and Harris 2004).

Sex-determination: Evolutionary analysis of sex-determination genes in bumblebees and related species indicated that all *Bombus* genomes share a duplicated copy of

feminizer (*fem*), named *fem 1* (Figure 4C). Compared to *fem*, *fem 1* shows a higher level of divergence among bumblebees (fem_{Bombus} $dN/dS = 0.24$; $fem\ 1_{\text{Bombus}}$ $dN/dS = 0.77$; Figure 4C). These ratios are close to the range observed for *Apis*, in which *fem* has evolved under purifying selection and the paralogous gene *complementary sex determiner* (*csd*) has evolved by neo-functionalization (Figure 4C) (Hasselmann, et al. 2008). A hypothesis branch-site testing framework (RELAX), identifies evidence for relaxation of selection in $fem\ 1_{Bombus}$ compared to fem_{Bombus} ($P<0.001$, $LR = 36.34$). Moreover, the spurious action of diversifying selection on branches was predominantly found in $fem\ 1_{Bombus}$ (Figure 4C). A mixed effect model of evolution (MEME) was applied to identify individual sites that were subject to episodic diversifying selection, and at least 15 sites ($p< 0.05$) were found to be under positive selection, with some being located in known motifs (Additional file 2: Figure S21). The results of these selection analyses suggest that both *fem* and *fem 1* contribute to the *Bombus* sex determination pathway. For the *transformer 2* (*tra-2*) gene, consistent amino acid changes between *Bombus* and *Apis* were found within the RNA recognition domain (Additional file 2: Figure S22), supporting a previous hypothesis of a regulatory modification between the two groups (Biewer, et al. 2015).

Discussion

Comparative analysis of multiple genomes in a phylogenetic framework substantially improves the precision and sensitivity of evolutionary inference and provides robust results identifying stable and dynamic features. In this study, we performed comparative analyses of genome structures and contents, as well as global and family-targeted gene evolutionary dynamics across the phylogeny of *Bombus,* using 17 annotated *de novo* assemblies and two previously published genomes.

Many attributes of bumblebee genomes are highly conserved across species. For example, overall genome size and genome structure, the number of protein-coding genes and non-coding RNAs, gene intron-exon structures, and the pattern of codon usage are all very similar across these 19 genomes. However, other aspects of genome biology are dynamically evolving. TEs are a major contributor to genome size variation (Figure 1) as well as a potential source of coding and regulatory sequences (Additional file 1: Table S10-12). Differential gene gain and loss also contribute to gene content variation across bumblebees and lead to lineage-specific gene repertoires (Figure 4 A; Additional file 2: Figure S20; Additional file 1: Table S17). Finally, for genes shared by all species, the action of positive selection is different across species (Additional file 1: Table S26; Additional file 1: Table S28; Additional file 1: Table S30; Additional file 1: Table S31), which can lead to gene functional divergence possibly reflecting key eco-ethological differences.

An exception to the otherwise overall conserved genome structure is the set of species in the subgenus *Psithyrus*. These bumblebees exhibit social parasitism; they do not have a worker caste, and it is not necessary for them to forage for nectar and pollen (Lhomme and Hines 2019). Originally, this subgenus was argued to be a separate genus due to distinct behavior and higher chromosome number, however subsequent phylogenetic analysis placed *Psithyrus* within the subgenus *Bombus* (Williams, et al. 2008). Here, based on a much larger genomic dataset, we confirm that species in the subgenus form a monophyletic group within the *Bombus* clade (Figure 1A). In addition, we show that, although *Psithyrus* species have an increased chromosome number, their genome sizes are within the range of those of the other bumblebees (Figure 1A), and their 25

chromosomes reflect a mix of fission, fusion, and retention of the 18 ancestral bumblebee chromosomes (Figure 2; Additional file 2: Figure S6). Chromosome rearrangements (e.g., fissions, fusions, and inversions) have been posited to play roles in speciation (Ayala and Coluzzi 2005), and thus may explain the diversification and social parasitic behavior of *Psithyrus*. In addition to genome structure variation, we identified a net loss of 11 odorant receptor genes in the common ancestor of *Psithyrus* species (Figure 4), which could be a cause or consequence of their socially parasitic behavior.

Bumblebee species exhibit different food preferences (Goulson and Darvill 2004; Sikora and Kelm 2012; Somme, et al. 2015), but the genetic basis underlying such variation is unknown. Like in other insects, smell and taste are used to distinguish different food sources (Kunze and Gumbert 2001; Ruedenauer, et al. 2015). In this study, we found out that genes involved in smell and taste perception are among the fastest evolving gene categories, both in copy number variation and in sequence divergence (Figure 3; Additional file 2: Figure S15; Additional file 1: Table S18-19). Therefore, the dynamic evolution of genes involved in smell and taste perception likely contribute to different food preferences, improved understanding of which could inform the use of new species in agricultural settings. Bumblebees exhibit rich morphology differences across species (Williams 1994) and they show species-specific responses to insecticides (Baron, et al. 2017). Chitin is a major component of the insect cuticle and peritrophic matrix, and chitin metabolic processes are related to morphogenesis, resistance to insecticides, and the tolerance of toxins in food (Barbehenn 2001; Erlandson, et al. 2019; Merzendorfer and Zimoch 2003; Zhu, et al. 2016). Genes related to chitin metabolism are also among the fastest evolving functional categories in bumblebees, both in copy number variation and in sequence divergence (Figure 3; Additional file 2: Figure S15; Additional file 1: Table S18-19). These variable patterns of chitin-related gene evolution potentially underlie observed differences in morphology and insecticide resistance, which could influence the suitability of different species for commercial use. Across bumblebee genomes the fastest evolving genes are also related to processes including protein glycosylation, methylation, proteolysis, and tRNA aminoacylation for protein translation (Figure 3; Additional file 2: Figure S15; Additional file 1: Table S18-19). Protein glycosylation is involved in multiple physiological processes including growth, development, circadian rhythms, immunity, and fertility (Walski, et al. 2017). tRNA aminoacylation for protein translation process are involved in response to the changing environment (Pan 2013). Some genes that are not among the fastest evolving categories—for example, immune and detoxification genes, which are involved in the interaction of bumblebees with external environments—show differential patterns of positive selection in subsets of species (Additional file 1: Table S28; Additional file 1: Table S30), which can lead to gene functional divergence. Taken together, identification of the fastest evolving genes and those showing patterns of differential positive selection reveals substantial genetic variation across bumblebees. Future experimental investigations will be required to determine how the identified genetic variation is

linked to specific differences in traits such as food preference, morphogenesis, insecticide and pathogen resistance, and the response to changing environments.

In addition to our discoveries regarding protein-coding genes, we found that TE-related sequences likely contribute to the variation of coding and regulatory repertoires (Figure 1; Additional file 1: Table S10-12). Compared with non-*Mendacibombus* bumblebees, *Mendacibombus* species have smaller genomes (Figure 1) and relatively narrow geographical distributions (Williams, et al. 2016). Considering TEs are the major determinant of genome size difference, with evidence that they were domesticated in bumblebee genomes, TEs may be implicated in the dispersal of non-*Mendacibombus* species across the globe, as they have been in other taxa (Baduel, et al. 2019; Casacuberta and González 2013; Schrader and Schmitz 2019).

More recent range expansions or contractions are driven, at least in part, by global climate change. To survive, bumblebees may have to move northward or to higher elevations as the climate warms (Kerr, et al. 2015; Soroye, et al. 2020). The sequenced genomes of species collected at high-elevation sites (> 4000 m) and others collected at low elevations (< 2000 m) (Figure 1) represent high quality genomic resources for investigating genes involved in high-elevation adaptation. We identified genes showing signs of positive selection in at least two subgenera of high-elevation species but not in any of the low-elevation species (Additional file 1: Table S31). These include genes putatively involved in eye development, muscle integrity maintenance, and metabolism, highlighting the importance of successful food-searching in high-elevation habitats where food is scarce. Exploring these further and identifying additional genomic features linked to life at high altitudes will help to understand differential successes of bumblebee species in a changing world.

Conclusions

We have produced highly complete and accurate genome assemblies of 17 bumblebee species, including representatives from all of the 15 subgenera of *Bombus*. Our genus-wide comparative analysis of bumblebee genomes revealed how genome structures, genome contents, and gene evolutionary dynamics vary across bumblebees, and identified genetic variations that may underlie species trait differences in foraging, diet and metabolism, morphology and insecticide resistance, immunity and detoxification, as well as adaptations for life at high altitudes. Our work provides genomic resources that capture genetic and phenotypic variation, which should advance our understanding of bumblebee success and help identify potential threats. These resources form a foundation for future research, including resequencing and population genomics studies for functional gene positioning and cloning, which will inform the use of bumblebees in agriculture, as well as the design of strategies to prevent the decline of this important group of pollinators.

Materials and Methods

Sample collection and DNA extraction

Criteria including phylogenetic position, biological trait, geographic distribution, and specimen availability were applied to select species for whole genome sequencing. A total of 17 bumblebee species were selected (Additional file 1: Table S1), which span all of the 15 subgenera in the simplified classification system for the genus *Bombus* (Williams, et al. 2008). Among these, two species (*B. superbus* and *B. waltoni*) are from the subgenus *Mendacibombus*, which is sister to all other *Bombus* lineages; four species (*B. superbus*, *B. waltoni*, *B. skorikovi* and *B. difficillimus*) were collected at extremely high elevations (> 4000 m above sea level); two species (*B. turneri* and *B. skorikovi*) exhibit social parasitism; and one species (*B. polaris*) is endemic to the Arctic. In addition, species traits (i.e. range size, tongue length, parasite incidence, and decline status) vary across the selected bumblebees (Arbetman, et al. 2017). Samples were collected in the summer of 2016, with location and elevation information summarized in Additional file 1: Table S1. Their identities were confirmed by DNA barcoding as described (Hebert, et al. 2004). Genomic DNA was extracted from each specimen using the Gentra Puregene Tissue Kit (Qiagen). The abdomens of each sample were removed before DNA extraction to avoid microbial contamination.

Genome sequencing and assembly

Genomic DNA purified from one single haploid drone of each species was used to generate one "fragment" library with an insert size of 400 or 450 bp using the NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (NEB, USA). The prepared fragment libraries were sequenced on an Illumina HiSeq 2500 platform with a read length of 250 bp to produce overlapping paired-end shotgun reads (2 × 250 bp), and the target sequencing coverage was 100-fold or more for each species. Genomic DNA purified from multiple specimens of each species was used to generate four "jump" libraries (insert sizes: 4 kb, 6 kb, 8 kb, and 10 kb) according to reported methods (Heavens, et al. 2015). The prepared jump libraries were sequenced on an Illumina HiSeq X Ten platform, and paired-end reads (2 × 150 bp) were generated, with a sequencing depth of at least 40-fold coverage for each jump library. The sequencing results of "fragment" and "jump" libraries are summarized in Additional file 1: Table S2.

For each species, the 250 bp overlapping paired-end shotgun reads from the fragment library were processed using the software Seqtk (https://github.com/lh3/seqtk) to randomly subsample read pairs to achieve the total sequence length equivalent to ~60-fold sequencing coverage, a coverage recommended by the assembler we used (https://software.broadinstitute.org/software/discovar/blog/). Then, the subsampled shotgun reads were assembled using the software DISCOVAR *de novo* (version 52488), which performs well in assembling insect genomes (Love, et al. 2016), to produce contiguous sequences (contigs) for each species. Finally, shotgun reads from jump libraries were used to scaffold the contigs using the software BESST (Version 2.2.6) (Sahlin, et al. 2014). The obtained genome assemblies were checked for DNA

contamination by searching against the NCBI non-redundant nucleotide database (Nt) using BLASTN (Camacho, et al. 2009), with an E-value cutoff of 1e-5.

To evaluate the quality and completeness of the genome assemblies, we compared genes present in the assemblies to a set of 4,415 universal single-copy orthologs (lineage dataset: hymenoptera_odb9) using the software BUSCO v3 (Waterhouse, et al. 2018).

Genome annotation

RNA extraction and sequencing. For each species (*B. superbus, B. waltoni, B. confusus, B. soroeensis, B. consobrinus, B. difficillimus, B. haemorrhoidalis, B. turneri, B. opulentus, B. picipes, B. ignitus, B. sibiricus, B. breviceps,* and *B. pyrosoma*), total RNA was isolated using the TRIzol reagent (Invitrogen, CA, USA) following the manufacturer's instructions. RNA integrity was evaluated on a 1.0 % agarose gel stained with ethidium bromide. After quantifying the concentration of RNA using a Qubit® 2.0 Fluorometer (Life Technologies, CA, USA), 3 μg of RNA from each species was used to prepare sequencing libraries using the NEBNext® UltraTM RNA Library Prep Kit for Illumina® (NEB, USA) following manufacturer's instructions. Library quality was assessed on the Agilent Bioanalyzer 2100 system. The prepared libraries were sequenced on the Illumina HiSeq X Ten platform, generating paired-end reads with a read length of 150 bp.

Protein-coding gene annotation. Annotation of protein-coding genes was based on *ab initio* gene predictions, transcript evidence, and homologous protein evidence, all of which were implemented in the MAKER computational pipeline (Cantarel, et al. 2008). Briefly, RNA-seq samples were assembled using Trinity (Haas, et al. 2013) with two different strategies using default parameters, *de novo* assembly and genome-guided assembly. Assembled transcripts were inspected by calculation of FPKM (fragments per kilobase of exon per million fragments mapped) expression values and removed if FPKM <1 and iso-percentage <3%. The filtered transcripts were imported into the PASA program (Haas, et al. 2003) for construction of comprehensive transcripts, as PASA is able to take advantage of the high sensitivity of reference-based assembly while leveraging the ability of *de novo* annotation to detect novel transcripts. The nearly "full-length" transcripts selected from PASA-assembled transcripts were imported to data training programs including SNAP (Korf 2004), GENEMARK (Lomsadze, et al. 2005) and AUGUSTUS (Stanke, et al. 2006). Afterwards, the MAKER pipeline was used to integrate multiple tiers of coding evidence and generate a comprehensive set of protein-coding genes.

The second round of MAKER was run to improve gene annotation. The predicted gene models with AED scores less than 0.2 were extracted for re-training using SNAP, GENEMARK, and AUGUSTUS. In addition, the RNA-seq reads were mapped to genomes using HiSAT2 and re-assembled using StringTie (Pertea, et al. 2016). The assembled RNA-seq transcripts, along with proteins from bees (superfamily Apoidea) that are available in NCBI GenBank (last accessed on 01/28/2018), were imported into the MAKER pipeline to generate gene models, followed by manual curation of key gene families.

Functional annotation of the obtained gene models

To obtain functional clues for the predicted gene models, protein sequences encoded by them were searched against the Uniprot-Swiss-Prot protein databases (last accessed on 01/28/2018) using the BLASTp algorithm implemented in BLAST suite v2.28 (Altschul, et al. 1990). In addition, protein domains and GO terms associated with gene models were identified by InterproScan-5 (Jones, et al. 2014).

To evaluate the quality and completeness of gene annotation, we compared protein sequences predicted from the genome assemblies to a set of 4,415 universal single-copy orthologs (lineage dataset: hymenoptera_odb9) using the software BUSCO v3 (Waterhouse, et al. 2018).

miRNA annotation

Hairpin sequences downloaded from miRBase (http://www.mirbase.org/) were aligned to each reference genome using BLASTN (Altschul, et al. 1990) with an e-value cut-off of 10-6. Results were further filtered based on alignment length (≥50nt) and sequence similarity (≥80%). Mature sequences from miRBase were then mapped against this set of selected BLASTN hits, using Patman (Prufer, et al. 2008) with parameters -g 0 -e 1 (no gaps, up to one mismatch). Only genomic hits where at least one mature microRNA could be mapped with these criteria were retained. These were treated as a set of putative homologous microRNA genes.

Small RNA reads of *B. terrestris* were mapped to these predicted homologous loci, with no gaps or mismatches allowed. Genomic loci with at least 10 mapped reads were then selected, showing coverage at both the 5' and 3' ends. The final set of high confidence microRNAs was obtained by selecting all loci with the expected hairpin secondary structure, as predicted by RNAfold from the ViennaRNA package (Hofacker 2009), as well as strong evidence of Drosha-Dicer processing from the (manually inspected) patterns of small-RNA read alignments.

tRNA annotation

All of the bumblebee genomes were screened with tRNAScan-SE (Lowe and Eddy 1997) to identify tRNA genes, with default parameters.

The prediction of lncRNAs: protein-coding potential for RNA transcripts was predicted using two algorithms, LGC version 1.0 (Wang, et al. 2019) and CPAT version 2.0.0 (Wang, et al. 2013). LGC could be used in a cross-species manner and the algorithm was applied directly to bumblebees, while CPAT requires high-quality training data to build a species-specific model. Considering bumblebees do not have enough high-quality "coding" and "non-coding" transcripts to build a model, the prebuilt fly model in CPAT was used. All the predictions were performed on a Linux platform. RNA transcripts were deemed to be non-coding if they were consistently predicted to be non-coding by both LGC and CPAT.

Gene synteny analysis

MCScanX (Wang, et al. 2012) was used to identify syntenic blocks, defined as regions with more than five collinear genes, between *B. terrestris*, a previously published bumblebee genome (Sadd, et al. 2015), and each of the newly sequenced bumblebees with default parameters to infer synteny contiguity.

***De novo* identification and annotation of transposable elements (TEs)**

Methods based on TE structure

LTR retrotransposons of the bumblebee genomes were *de novo* identified and annotated by LTRharvest and LTRdigest (Ellinghaus, et al. 2008; Steinbiss, et al. 2009). The identified LTR retrotransposons were further classified with the PASTEC module of the REPET package (Hoede, et al. 2014). When identifying LTR retrotransposons, TSD length was set to 4-6 bp and the minimum similarity of LTRs was set to 85%; the four-nucleotide termini of each LTR retrotransposon was set as TG…CA. LTR length was set to 100-6000 bp. For the post-processing of LTRdigest, pptlength was set to 10-30 bp, pbsoffset to 0-5 bp, and trans to Dm-tRNAs.fa. pHMMs were used to define protein domains taken from the Pfam database. Non-LTR retrotransposons of the bumblebee genomes were identified and characterized using MGEScan-non-LTR, with default parameters (Rho and Tang 2009).

DNA transposons were identified by TBLASTN of known DNA transposase sequences that are available in Repbase (https://www.girinst.org/repbase/) against the bumblebee genome sequences. All regions that produced significant hits (E-values <1E-10) were excised with 3 kb of flanking regions. The terminal inverted repeats of a DNA transposon were identified through a self-alignment of the excised sequence using NCBI-BLAST 2.

Methods based on the repetitive nature of TEs

RepeatScout (Price, et al. 2005) was used to *de novo* identify repetitive sequences from bumblebee genomes, with default parameters. The obtained consensus sequences were classified by the PASTEC module of the REPET package (Hoede, et al. 2014). All of the repetitive sequences were classified into Class I (retrotransposons), Class II (DNA transposons), Potential Host Genes, SSR (Simple sequence repeats) and “noCat” (which means no classification was found).

TE landscapes in the bumblebee genomes

First, CD-HIT-EST (version 4.6.6) (Li and Godzik 2006) was used to parse TE sequences that were *de novo* identified based on structure and repetitive nature with a sequence identity threshold of 0.9 (other parameters as default) to reduce TE redundancy for each bumblebee species. Then, the remaining TE sequences from all the bumblebee species were combined to produce a comprehensive TE library. Using this repeat library, each bumblebee genome was analyzed with RepeatMasker (http://www.repeatmasker.org) to yield a comprehensive summary of the TE landscape in each species using Cross_Match as the search engine (other parameters as default). The annotation files produced by RepeatMasker were processed by in-house scripts to eliminate redundancy. Refined annotation files were used to determine the TE diversity and abundance within each species. Tandem repeats in each genome were identified by Tandem Repeat Finder (Benson 1999), implemented in RepeatMasker.

TEs proliferated after the divergence of *Mendacibombus* from the other subgenera

The subgenus *Mendacibombus* forms the sister group to all of the other extant bumblebees, diverging near the Eocene-Oligocene boundary approximately 34 million years ago (Cameron, et al. 2007; Hines 2008; Williams and Paul 1985). If a

TE is present in one non-*Mendacibombus* species, but is absent at the orthologous positions in both *Mendacibombus* species (*B. superbus* and *B. waltoni*), then the TE is inferred to have transposed sometime after the divergence of the species from *Mendacibombus*. To identify such TEs in each of the non-*Mendacibombus* species, first, pairwise whole-genome alignments between the target species and *B. superbus* were performed using the software LASTZ (Harris 2007). Then, based on the whole genome alignment results, TE insertion scanner (https://github.com/Adamtaranto/TE-insertion-scanner) was used to identify "alignment gaps" showing signatures of TE insertions in the genome of the target species, with "--maxInsert 50000 --minIdent 85 --minInsert 80" choices (other parameters set as default). Secondly, 200 bp of sequence flanking the identified TE-like insertion on either side were extracted from the genomic sequences of the target bumblebee species and combined into one sequence of 400 bp. Then, the flanking sequences were used as queries in BLASTn searches against the genomic sequence of *B. waltoni*, with an e-value cutoff of 1e-10. Hits spanning both sides of the TE-like insertion with a minimal length of 350 bp were considered as empty sites in *B. waltoni* genome. Finally, TE-like sequences that have identifiable orthologous empty sites in both of the two *Mendacibombus* species were RepeatMasked by the comprehensive TE library of bumblebees to confirm their TE identity.

The age distribution of TE families in bumblebees

The consensus sequence of each TE family was constructed using RepeatScout (Price, et al. 2005) for each of the 19 bumblebee species; this consensus represents the TE family's master gene (i.e. ancestral sequence). The obtained consensus sequences were used to produce a species-specific TE library. Using these libraries, each genome was masked with RepeatMasker. Percent divergences from consensus sequences reported by RepeatMasker were converted to nucleotide distance measures using the Jukes-Cantor formula to correct for multiple hits. To increase accuracy, analyses were limited to TE elements ≥80% identical to their respective consensus sequences, with a minimum length of 80 bp. Results were pooled into bins of single unit distances and represent summaries of TE class proliferation history. Because TEs evolve neutrally following insertion, the age of individual TEs can be approximated by measuring the sequence divergence from the ancestral consensus sequence and by applying a neutral substitution rate of 3.6×10^{-9} for bumblebee (Liu, et al. 2017).

The genomic distribution of TEs in bumblebees

The genomic coordinates of TEs in each species were compared with the coordinates of protein-coding genes in the same species to identify TEs that resided within or near predicted genes. Only when there were > 50 bp of overlap between a TE and predicted CDS was a TE considered to be overlapping with a coding region. In *B. terrestris*, the coordinates of TEs, excluding those found in coding regions, were also compared with the coordinates of open chromatin regions detected by ATAC-seq (Zhao, et al. 2019) to identify TEs that may serve as regulatory sequences. Orthologous groups containing genes whose coding regions have TE-derived sequences were extracted, along with their *dN/dS* values (see Molecular evolution

analysis on gene functional categories section) to check their *dN/dS* ratios to determine if they are under selective constraint.

Orthology delineation across *Apis* and *Bombus*

The locally installed OrthoDB pipeline (http://www.orthodb.org/software; Kriventseva et al., 2015) was employed to define orthologous groups for proteins coming from 19 bumblebees and 4 honeybees. In addition to the 17 newly sequenced bumblebees from this study, the following previously annotated gene sets were downloaded: *B. terrestris* (GenBank assembly: Bter_1.0), *B. impatiens* (GenBank assembly: BIMP_2.0), *Apis mellifera* (GenBank assembly: Amel_4.5), *Apis cerana* (GenBank assembly: ACSNU-2.0), *Apis florea* (GenBank assembly: Aflo_1.0), and *Apis dorsata* (GenBank assembly: *Apis dorsata* 1.3). Only the longest isoform of each gene was used in orthology delineation. The orthoMCL program (Li, et al. 2003) was applied to the same protein dataset to confirm the results of the OrthoDB pipeline on lineage- and species-specific genes, and only genes determined as lineage- or species-specific by both programs were used for downstream analysis. In order to characterize the function of *Bombus*-specific genes, genes from *B. terrestris* that are *Bombus*-specific were selected. The GO annotations of *Bombus*-specific genes were assigned by InterproScan-5 (Jones, et al. 2014) and visualized on the WEGO website (http://wego.genomics.org.cn/; gene level 4) (Ye, et al. 2006).

To construct the phylogeny for these 23 species (19 bumblebees and 4 honeybees), universal single-copy orthologs delineated by the OrthoDB pipeline were isolated, and 3,617 single-copy orthologs were identified. Protein sequences from each of those universal single-copy orthologs were aligned with the software MAFFT (Katoh, et al. 2002), followed by alignment trimming with BMGE (Criscuolo and Gribaldo 2010). Trimmed alignments were concatenated for each species, respectively, resulting in 23 long super-sequences. The super-alignment contained 2,008,306 amino acids with 222,460 distinct alignment patterns. IQTree version 2.0 (Minh, et al. 2020b) was used to construct a maximum likelihood concatenated tree with the ultrafast bootstrap method (Hoang, et al. 2018). The best-fitting amino acid substitution model for each partition was selected by automatically by IQTree's internal implementation of ModelFinder (Kalyaanamoorthy, et al. 2017). A time calibrated, ultrametric tree was produced by using a non-parametric rate smoothing approach (Sanderson 2003) along with a fossil calibration range of 65 My to 125 My for the divergence of *Apis* and *Bombus* (Hines 2008). To assess phylogenetic discordance among loci, gene trees for each single-copy orthologous group were also reconstructed with IQTree (Additional file 1: Table S5)(Minh, et al. 2020b). Of the 3,617 gene trees, 3,530 could confidently be rooted by the outgroup genus *Apis* to count topologies (Additional file 1: Table S6). Rooting was performed using Newick Utilities (Junier and Zdobnov 2010). Gene and site concordance factors (CF) were then calculated for each node in the species tree as implemented in IQTree (Minh, et al. 2020a).

The quartet-based species tree reconstruction program ASTRAL (Zhang, et al. 2018), which can account for ILS, was also used for building the species phylogeny. The ggtree R package was used to visualize trees (Yu, et al. 2017).

Estimate of ancestral genome sizes

The genome assemblies produced in this study were highly complete (Additional file 2: Figure S1), and genome assembly sizes do not correlate with assembly contiguity ($p = 0.973$; Additional file 2: Figure S23). Thus, smaller genome size estimates are unlikely to be artifacts of incomplete genome assembly, and quality control during assembly ensured that larger genomes were not due to extrinsic DNA contamination. Therefore, the genome assembly sizes should reflect true differences across bumblebees. Genome assembly sizes of the 19 sequenced bumblebees and four honeybees were obtained from the current study and published genome assemblies: *B. terrestris* (GenBank assembly: Bter_1.0), *B. impatiens* (GenBank assembly: BIMP_2.0), *Apis mellifera* (GenBank assembly: Amel_4.5), *Apis cerana* (GenBank assembly: ACSNU-2.0), *Apis florea* (GenBank assembly: Aflo_1.1), and *Apis dorsata* (GenBank assembly: *Apis dorsata* 1.3). Genome sizes were mapped onto the phylogenetic tree estimated in this study (Figure 1A), and ancestral genome sizes of bumblebees were estimated using parsimony ancestral state reconstruction in Mesquite 3.51 (http://www.mesquiteproject.org), with honeybee genome sizes serving as the outgroup.

Hi-C library construction, sequencing, and assembly

For *B. turneri*, library preparation was performed by Annoroad Gene Technology (http://en.annoroad.com) and mainly followed a protocol described previously (Belton, et al. 2012). Briefly, thorax muscles of wild-caught males were cross-linked by 2% formaldehyde solution at room temperature for 20 mins, and 2.5 M glycine was added to quench the crosslinking reaction. After grinding with liquid nitrogen, homogenized tissues were resuspended in 25 ml of extraction buffer I (10 mM Tris-HCl [pH 8.0], 5 mM β-mercaptoethanol, 0.4 M sucrose, 10 mM MgCl2, 0.1 mM phenylmethylsulfonyl fluoride [PMSF], and 1x protease inhibitor [Roche]), then filtered through miracloth (Calbiochem). The filtrate was centrifuged at 3,500g at 4°C for 20 min. The pellet was resuspended in 1 ml of extraction II (10 mM Tris-HCl [pH 8], 0.25 M sucrose, 10 mM MgCl2, 1% Triton X-100, 5 mM β-mercaptoethanol, 0.1 mM PMSF, and 1x protease inhibitor) and then centrifuged at 18,400g and 4 °C for 10 min. The pellet was resuspended in 300 µl of extraction buffer III (10 mM Tris-HCl, [pH 8.0], 1.7 M sucrose, 0.15% Triton X-100, 2 mM MgCl2, 5 mM β-mercaptoethanol, 0.1 mM PMSF, and 1 x protease inhibitor) and loaded on top of an equal amount of extraction buffer III, then centrifuged at 18,400g for 10 min. The supernatant was discarded and the pellet was washed twice by resuspending it in 500 µl of ice-cold 1x CutSmart buffer, followed by centrifuging the sample for 5 min at 2,500g. The nuclei were washed by 0.5 ml of 1 x restriction enzyme buffer and transferred to a safe-lock tube. Next, the chromatin was solubilized with dilute SDS and incubated at 65 °C for 10 min. After quenching the SDS with Triton X-100, overnight digestion was applied with a four-cutter restriction enzyme (400 units of MboI) at 37 °C on a rocking platform. The flowing steps include marking the DNA ends with biotin-14-dCTP and performing blunt-end ligation of crosslinked fragments. The proximal chromatin DNA was re-ligated by ligation enzyme. The nuclear complexes were reverse-crosslinked by incubating with proteinase K at 65 °C. DNA was purified by phenol–chloroform extraction. Biotin-C was removed from

non-ligated fragment ends using T4 DNA polymerase. Fragments were sheared to a size of 100–500 bp by sonication. The fragment ends were repaired by the mixture of T4 DNA polymerase, T4 polynucleotide kinase, and Klenow DNA polymerase. Biotin-labeled Hi-C samples were specifically enriched using streptavidin magnetic beads. A-tailing of the fragment ends were added by Klenow (exo-) and Illumina paired-end sequencing adapters were added by ligation mix. Finally, Hi-C sequencing libraries were amplified by PCR (12-14 cycles) and sequenced on the Illumina HiSeq X Ten platform, generating paired-end reads (2 × 150 bp). The Juicer tool (Durand, et al. 2016) was applied to map Hi-C reads against the contig sequences of *B. turneri* using the BWA algorithm (Heng, et al. 2010) with default parameters. Mapped reads with MAPQ quality scores ≥ 30 were chosen for the next analysis. Then, the 3D-DNA pipeline (Dudchenko, et al. 2017) was applied to assemble the scaffold sequences to the chromosome level.

For *B. ignitus*, *B. pyrosoma*, *B. breviceps*, and *B. haemorrhoidalis*, the *in situ* Digestion-ligation-only Hi-C protocol was employed to generate Hi-C reads as described (Lin, et al. 2018). In brief, for each species, brain tissue of wild-caught workers was ground into homogenate. Treated the samples and filtered the precipitated cells. Cells were double cross-linked with formaldehyde with EGS (Thermo) and 1% formaldehyde (Sigma). After that, the remaining formaldehyde was sequestered with glycine. The cross-linked cells were subsequently lysed in lysis buffer and incubated at 50 °C for 5min, placed on ice immediately. After incubation, the nuclei were digested by MseI (NEB, 100 units/μl). After restriction enzyme digestion, MseI biotin linkers were ligated to the digested chromatin respectively. Made the nuclei fragment-end phosphorylation. Next, added T4 DNA ligase (Thermo) to reaction complexes. Ligation was performed at 20 °C for 2h with rotation at 15 r.p.m. Then, purifying the proximity ligation DNA. The purified products were digested by MmeI at 37 °C for 1 h. The digested DNA sample was subjected to electrophoresis in native PAGE gels and the specific 80-bp DLO Hi-C DNA fragments were excised and purified. Next, Illumina sequencing adaptors were ligated to the 80-bp DLO Hi-C DNA fragments. After biotin incubation, the ligated DNA fragments were used as template and amplified by PCR (fewer than 13 cycle) to construct the Illumina sequencing libraries.

Hi-C sequencing libraries were sequenced on the Illumina HiSeq X Ten platform, generating 150 bp reads. The length of the DNA constructs in the DLO Hi-C library is between 78 and 82 bp. The length of a full linker is 40 bp, and the lengths of the target DNA sequences on each side of the linker are 19-21 bp. A Java program was used to exclude the linker parts from the reads and the target DNA fragments were used for downstream analysis. The Juicer tool (Durand, et al. 2016) was applied to map obtained target sequences against the scaffold sequences of each species using the BWA algorithm (Heng, et al. 2010), selecting the ALN parameter (other parameters as default). Mapped reads with MAPQ quality scores ≥ 30 were chosen for the next analysis. Then, the 3D-DNA pipeline (Dudchenko, et al. 2017) was applied to assemble the scaffold sequences to the chromosome level.

The coordinates of genes within scaffold sequences were converted into coordinates on chromosome sequences for those five species.

Macrosynteny search and visualization

First, the longest CDS for each gene, along with their coordinates, were prepared for the bumblebee species with chromosome-level assemblies (*B. ignitus*, *B. pyrosoma*, *B. breviceps*, *B. haemorrhoidalis*, *B. terrestris* and *B. turneri*). Then, pairwise comparisons were performed between *B. turneri* and each of the other species using MCscan in the JCVI tool kit (https://github.com/tanghaibao/jcvi; last accessed Dec 25, 2019) (Wang, et al. 2012) to identify and visualize macrosynteny.

Evaluation of chromosomal evolution rates

Orthologous genes and their coordinates on chromosomes were used as anchors to evaluate rates of chromosomal evolution. Two sets of orthologous genes for each pair of species were grouped together to form a standard input for the GRIMM-Synteny program v. 2.02 (Tesler 2002). The genome of *B. terrestris* was used as a reference for pairwise comparisons with other species genomes. Chromosomes of different species with similar sets of genes were named chromosomal elements. The GRIMM-Synteny program was run with default settings and the rearrangement distances (the number of conserved synteny blocks and inversions) were summarized.

Global gene family evolution analysis

In order to identify rapidly evolving gene families within *Bombus*, protein sequences from the following species were used: *B. superbus, B. confusus, B. soroeensis, B. consobrinus, B. difficillimus, B. haemorrhoidalis, B. turneri, B. opulentus, B. picipes, B. ignitus, B. polaris, B. cullumanus, B. sibiricus, B. breviceps, and B. pyrosoma* (one species per subgenus was selected to avoid over-sampling in any subgenus). To ensure that each gene was counted only once, only the longest isoform of each gene in each species was used. An all-vs-all BLAST (Altschul, et al. 1997) search was then performed on these filtered sequences. The resulting e-values from the search were used as the main clustering criterion for the MCL program to group proteins into gene families (Enright and J. 2002). This resulted in 24,137 clusters. All clusters only present in a single species or not present at the root of the tree were then removed, resulting in 13,828 gene families. A time calibrated, ultrametric tree (Additional file 2: Figure S11) was built by taking the inferred *Bombus* phylogeny and using a non-parametric rate smoothing approach (Sanderson 2003) along with a fossil calibration range of 65 My to 125 My for the divergence of *Apis* and *Bombus* (Hines 2008).

With the gene family data and ultrametric phylogeny as input, gene gain and loss rates (λ) were estimated with CAFE v3.0 (Han, et al. 2013). This version of CAFE is able to estimate the amount of assembly and annotation error (ε) present in the input data using a distribution across the observed gene family counts and a pseudo-likelihood search. CAFE is then able to correct for this error and obtain a more accurate estimate of λ. The resulting ε value was about 0.05, which implies that 5% of gene families have observed counts that are not equal to their true counts. After correcting for this error rate, $\lambda = 0.0036$. Using the estimated λ value, CAFE infers ancestral gene counts and calculates p-values across the tree for each family to assess

the significance of any gene family changes along a given branch. Those branches with low p-values are inferred to be rapidly evolving. A Fisher's exact test was performed on GO terms for genes in families that are rapidly evolving on any lineage vs. all other families, with a false discovery rate of 0.01.

Protein domain variation across bumblebees

Predicted protein sequences were analyzed by InterproScan-5 (Jones, et al. 2014) to identify InterPro domains in each bumblebee species. InterPro domain annotations across the 19 bumblebee species were used to identify protein domains exhibiting the highest variation in gene counts across bumblebees. A crude measure that highlights such variation in copy-number was computed as the standard deviation divided by the mean of the bumblebee gene counts matching a particular InterPro domain. Results were filtered to focus on abundant domains, which have more than 200 genes in total and more than five genes in each bumblebee species.

Molecular evolution analysis on gene functional categories

Orthology delineation across bumblebees: In addition to the 17 newly sequenced bumblebees from this study, we downloaded the two previously annotated gene sets for *B. terrestris* and *B. impatiens* from Ensembl (http://metazoa.ensembl.org/index.html). Only the longest isoform of each gene was used for downstream analysis. Protein sequences from the 19 bumblebees were used to delineate orthologous groups by locally installed OrthoDB software (OrthoDB_soft_2.4.4) (http://www.orthodb.org/software).

Assignment of functional categories to each orthologous group: GO term(s) and InterPro domain(s) associated with each gene of the orthologous group were identified by InterproScan-5 (Jones, et al. 2014). A GO term or InterPro domain was assigned to this orthologous group if more than 60% of the genes in it were assigned this GO term or InterPro domain by InterproScan-5.

Evolutionary rate (amino acid sequence divergence) estimation for each orthologous group: Evolutionary rates were computed for each orthologous group as the average of inter-species identities normalized to the average identity of all inter-species best reciprocal hits, computed from pairwise Smith-Waterman alignments of protein sequences. The 'evolrate' program of the OrthoDB_soft_2.4.4 package was used to obtain these rates.

***dN/dS* ratio estimation for each orthologous group**: To avoid biases related to duplication among lineages and out-paralog genes, only universal single-copy orthologous groups (scOGs) were used to estimate *dN/dS* ratios. Protein sequences of scOGs were aligned by MAFFT (Katoh, et al. 2002) and then used to inform CDS alignments to generate DNA codon alignments with the codon-aware PAL2NAL program (Suyama, et al. 2006). Next, the aligned CDSs were trimmed by Gblocks (Talavera and Castresana 2007), with "-t c" and other parameters as default. After trimming, only orthologs consisting of aligned sequences from all species with a minimum of 150 bp and less than 20% Ns were retained for downstream analysis, which are available on-line (ftp://download.big.ac.cn/bumblebee/bumblebee-single-copy-orthologs.tar.gz). Then, based on trimmed alignments, Maximum Likelihood trees were constructed for each of the orthologous groups using RAxML-NG

(Kozlov, et al. 2019). Finally, PAML (Yang 2007) was used to calculate the *dN/dS* ratio for each orthologous group using its respective phylogenetic tree (codeml model=0, NSsites=0, ncatG=1).

Enrichment analysis of the slowest and fastest evolving genes: Assignment of GO terms and InterPro domains was biased towards slower-evolving, well-conserved genes (Additional file 2: Figure S14), so the fastest evolving genes are less likely to be functionally annotated. Comparing the top enriched functional categories in the slowest and fastest subsets of genes could complement the GO and InterPro analyses described above. Orthologous groups with evolutionary rates and *dN/dS* ratios less than the 20th percentile or greater than the 80th percentile were selected to represent the slowest and fastest gene sets, respectively (Additional file 2: Figure S16). Enrichment tests on GO Biological Processes and Molecular Functions were performed using Bioconductor's GOstats hypergeometric test (Falcon and Gentleman 2007) and with the topGO (http://www.bioconductor.org/packages/release/bioc/html/topGO.html) implementations of the classic Fisher and the weighted Fisher tests. The background gene sets in each case were genes from all 19 bumblebee genomes that were classified into any orthologous group and were annotated with Biological Process or Molecular Function GO-terms. The results were combined using a conservative strategy: terms must appear significant with a p-value <0.05 for all three enrichment tests, and there must be more than five genes in the test set. Complementary enrichment analyses using topGO's implementation of the Kolmogorov–Smirnov (KS) were performed using evolutionary feature metrics: evolutionary rate (as above); universality (the proportion of species with genes in each orthologous group); and three copy-number metrics (average copy-number, copy-number variation, and proportion of species with duplicates). Only Biological Process terms associated with at least 10 orthologous groups were assessed. The KS test uses the score distributions directly without having to specify any top or bottom cut-off as described above for the classic tests with the 20th and 80th percentiles. Results are presented for terms showing significantly higher or significantly lower score distributions (Additional file 1: Table S14; Additional file 1: Table S19).

Detection of positive selection signatures

(1) Single-copy orthologous groups search: orthologous groups containing focal genes, along with their *dN/dS* values, were extracted from the **Molecular evolution analysis on gene functional categories** section. To avoid biases related to duplication among lineages and out-paralog genes, only universal single-copy orthologous groups were kept for downstream analysis.

(2) Multiple sequence alignment and *de novo* gene tree construction: The multiple alignment and Maximum Likelihood tree of each ortholog were taken from the **Molecular evolution analysis on gene functional categories** section.

(3) aBSREL analysis: For each ortholog, signatures of positive diversifying selection were searched using the aBSREL algorithm (https://www.datamonkey.org), with the respective multiple sequence alignment and Maximum Likelihood tree. Branches with test p-values < 0.05 were considered to be under selection.

Intron evolution

Orthologous groups delineated across 19 bumblebees and one honeybee (*A. mellifera*) (deduced from the **Orthology delineation across *Apis* and *Bombus*** section) were examined to select a total of 8,672 with near-universal single-copy orthologue distributions: requiring no more than two species with no orthologues and no more than two species with multi-copy orthologues. These were further filtered to exclude groups with genes for which annotation features did not match the protein sequence and groups where the orthologues from five or more of the 20 species were single-coding-exon genes (i.e. no introns), leaving 7,394 groups for the analysis. The protein sequences of the orthologues for each group were FASTA formatted with header information containing intron/exon data required for analysis with Malin (Csűros 2008). Protein sequences for each group were aligned with MAFFT v7.310 (Kazutaka and Standley 2013) using the '--auto' option. The resulting alignments were then processed (two rounds of re-alignment) by the IntronAlignment tool from the Malin suite with option '-matrix blosum62 -rep 2'. The species tree and alignments were loaded into the Malin analysis tool and reliable intron sites were defined as having at least five non-gap amino acid positions in the alignment before and after the site and unambiguous characters in at least 18 of the 20 species. This resulted in a total of 45,804 sites for the analysis which was performed using the Bootstrap Posterior Probability (BPP) approach of Malin, using rate models computed from the default starting model with default optimization parameters and with one gain and one loss level.

Stop codon readthrough analysis

Whole genome alignments: Before multiple whole genome alignments, repetitive regions of the 19 bumblebee and 4 honeybee (*Apis mellifera*, *Apis cerana*, *Apis florea*, and *Apis dorsata*) genome assemblies were first masked to reduce the total number of potential genomic anchors formed by the many matches that occur among regions of repetitive DNA. For whole genome alignments of the 23 bees, Cactus (Paten, et al. 2011), a reference-free whole genome aligner, was used. The phylogeny of 23-species estimated in this study (Figure 1A), with branch lengths reflecting neutral substitutions per site, was used as the guide tree.

Stop codon readthrough analysis: Annotation version GCF_000214255.1 for *B. terrestris*, obtained from NCBI, was used. The phylogeny is the 23-species maximum likelihood phylogeny estimated in this study. PhyloCSF (Lin, et al. 2011) was run on the region between the annotated stop codon ("first stop codon") and the next in-frame stop codon ("second stop codon") referred to as the "second open reading frame (ORF)", excluding both the first and second stop codons, of all annotated transcripts whose coding region ends in a stop codon, grouping together sets of transcripts having the same second ORF. For transcripts lacking an annotated 3'UTR, or for which the 3'UTR does not extend up to the second stop codon, the transcript was extened along the DNA strand without splicing. PhyloCSF was run using the default "mle" strategy and "bls" option, using the 12flies parameters but substituting the 23-bees tree. PhyloCSF computes a log-likelihood of an alignment under coding and non-coding models of evolution. The model assumes independence of codons

given that the region is coding or non-coding. However, scores of neighboring codons are not independent. To correct for that, PhyloCSF-Ψ (Lin, et al. 2011) calculates a log-likelihood of length-dependent normal distributions trained on actual coding and non-coding regions of various lengths. Coefficients for PhyloCSF-Ψ were trained using coding regions at the ends of coding ORFs and non-coding regions at the starts of third ORFs, as described in (Jungreis, et al. 2016). The coefficients we obtained for *B. terrestris* were:

$\mu_C = 0.678782322375$, $A_C = 8.09766004622$, $B_C = 0.783878652717$

$\mu_N = -6.80739917655$, $A_N = 9.51882863955$, $B_N = 0.664609908575$

Both raw PhyloCSF scores and PhyloCSF-Ψ scores are reported in units of decibans.

The 851 candidate readthrough stop codons in 817 genes were those satisfying all of the following conditions: (i) The second ORF is at least 10 codons long. (ii) PhyloCSF-Ψ > 0. (iii) The phylogenetic branch length of aligned species is more than 60% of the branch length of the full tree (enough to assure at least one *Apis* species is included). (iv) Species comprising at least 90% of the tree have the same first stop codon as *B. terrestris* (the *Drosophila* and *Anopheles* studies had found that readthrough stop codons are generally perfectly conserved). (v) Species comprising at least 60% of the tree have some stop codon aligned to the second stop codon. (vi) For second ORFs that overlap an annotated coding region on the same strand in the same reading frame, or on the opposite strand in the frame having the same third codon position (the "antisense" frame), the non-overlapping portion was required to be at least 10 codons long and have a positive PhyloCSF-Ψ score, as well as satisfying the branch length requirements described above.

To estimate the false discovery rate among our candidates, enrichment of the TGA stop codon with 3' base C was used, which is known to be the "leakiest" 4-base stop codon context (Bonetti, et al. 1995) and is highly enriched among readthrough stop codons (Jungreis, et al. 2011). Of the 851 stop codons in the list, 172 (20.2%) have the TGA-C context, whereas of the 8059 annotated stop codons for which the second ORF has negative PhyloCSF-Ψ score and are thus unlikely to be readthrough, only 280 (3.5%) have the TGA-C context. Among the readthrough stop codons previously reported in *Drosophila* 32.2% had the TGA-C context (Jungreis, et al. 2011). If a similar fraction holds in *Bombus*, the number of actual readthrough stops codons among the 851 would be approximately $(172 - 3.5\% \times 851) / (.322 - .035) = 496$. Even if as many as 50% of readthrough stop codons in *Bombus* use TGA-C, a similar calculation provides a conservative estimate that the list includes 306 readthrough transcripts. Among the 200 of the candidates with highest PhyloCSF-Ψ score, 72 have TGA-C stop context, so a similar calculation conservatively estimates 140 readthrough transcripts among these 200 candidates, for a false discovery rate of no more than 30%.

Codon usage bias analysis

Codon usage bias, the preferential use of specific synonymous codons, is a pattern maintained by mutation–selection–drift balance. The selection is linked to the efficiency and/or accuracy of translation. The selective effect of codon usage is only slightly advantageous and consequently selection's efficiency depends on population

size (Subramanian 2008; Vicario, et al. 2007); species with larger population sizes have more efficient selection for codon usage bias. Within the genomes, strength of selection could vary based on the stage of development when the genes are mainly translated (Vicario, et al. 2007). To determine the evolutionary forces affecting codon usage bias across bumblebees, a set of universal orthologous protein-coding genes was used (delineated in the **Molecular evolution analysis on gene functional categories** section). A total of 3,521 genes, which are present in all 19 species and have at least 50 unambiguous codons (no N or other ambiguity letters), were used for codon bias analysis. Candidate optimal codons were defined by examining the correlation between overall gene codon usage bias and the preference of use of a single codon as performed previously (Vicario, et al. 2007). As an estimator of overall codon usage bias, the Effective Number of Codons (ENC) was used, which was estimated by using the exponential of the sum of Shannon entropy of each codon family frequency set. As an estimator of preference for a single codon, the relative synonymous codon usage (RSCU) was used.

Gene family evolution analysis of chemosensory genes

To detect the putative chemosensory genes of the three major gene families □ odorant receptors (ORs), gustatory receptors (GRs) and ionotropic receptors (IRs) □ from the 17 newly sequenced and the *B. impatiens* genomes, TBLASTN searches (with 1e-5 as the e-value cutoff) (Gertz, et al. 2006; Karpe, et al. 2016) were performed using the protein sequences of *A. mellifera* (Robertson and Wanner 2006) and *B. terrestris* (Sadd, et al. 2015) as queries. Putative chemosensory gene-containing regions were extracted from each genome to predict gene models using the protein2genome module of Exonerate v2.2.0 (Slater and Birney 2005). These putative gene-containing regions were separately re-examined if there were no good hits based on Exonerate. Candidate chemosensory genes were further manually refined and checked for the characteristic domains of ORs (IPR004117), GRs (IPR009318 or IPR013604), or IRs (IPR019594 or IPR001320) in their encoded protein sequences using InterProScan v5.27-66.0 (Jones, et al. 2014; Zhou, et al. 2015; Zhou, et al. 2012). Partial sequences were completed with the nearest START and/or STOP codons wherever possible. Probable amino acid sequences of pseudogenes, which were identified using in-frame STOP codons or frameshifts, were determined from their predicted coding regions, and the letter "X" was used to represent STOP codons and frameshifts. The letter "Z" denotes unknown amino acids. The same procedure was repeated, using newly identified chemosensory genes as queries, until no additional genes were found. Gene names were assigned following the closest homologue of *B. terrestris*. When there were two or more gene copies in one analyzed species but a single-copy in *B. terrestris*, candidate gene names were suffixed with a, b, c, and so on. For ORs and GRs, genes encoding intact proteins with a length >= 350 amino acids were kept for downstream analysis.

Multiple alignments of the available bumblebee chemosensory genes were generated using MAFFT v7.407 ("E-INS-i strategy) (Kazutaka and Standley 2013), poorly aligned regions in the alignments were filtered using TrimAl v1.4 ("automated1" option) (Capellagutierrez, et al. 2009), and maximum-likelihood phylogenetic trees

were estimated using RAxML v8.2.11 (with the "PROTCATJTTF" model and 100 bootstrap replicates) (Stamatakis 2014). To estimate the numbers of gains and losses of chemosensory genes, we used maximum-likelihood-based and parsimony-based approaches, respectively; all genes of each chemoreceptor family were used as input for CAFE v4.2 (De Bie, et al. 2006) with default settings, and gene trees were reconciled with species tree using Notung v2.9.1 (Chen, et al. 2000).

Signatures of positive selection were searched for OR, GR and IR genes as described in **Detection of positive selection signatures** section.

Evolution of genes involved in detoxification

Glutathione-S-transferases (GSTs), carboxyl/cholinesterases (CCEs), and cytochrome P450 monooxygenases (P450s) are involved in the detoxification of xenobiotics. To identify detoxication genes in the newly sequenced bumblebees, annotated P450, GST, and CCE protein sequences of *B. terrestris*, *A. mellifera*, and *D. melanogaster* were used as queries to search against the predicted protein sequences from each genome using BLASTp (Altschul, et al. 1990). If certain genes appeared to be missing, TBLASTn was used as in annotating chemosensory genes. All of the identified detoxication genes were further checked for the presence of their characteristic domains to confirm their identity (GST [IPR004045 and IPR010987], P450 [IPR001128], and CCE [IPR002018]).

Signatures of positive diversifying selection were searched for each category of detoxication genes as described in **Detection of positive selection signatures** section.

Identification and characterization of immune genes

To identify immune-related genes in the newly sequenced bumblebees, annotated immune genes of *B. terrestris* and *A. mellifera* were used as queries to search against the predicted protein sequences from each genome using BLASTp (Altschul, et al. 1990). If certain genes appeared to be missing, TBLASTn was used as in annotating chemosensation genes.

Immune genes were classified into three broad functional categories — "recognition," "signaling," and "effector" — based on previous reports (Barribeau, et al. 2015; Evans, et al. 2006; Neafsey, et al. 2015; Sackton, et al. 2007; Waterhouse, et al. 2020). Specifically, the recognition class includes SCR (scavenger receptors), GNBP (gram-negative binding proteins), PGRP (peptidoglycan recognition proteins), and GALE (galectins). The signaling class includes TOLL (toll-like receptors), JAKSTAT (Jak/Stat pathway members), IMDPATH (Imd pathway members), CLIP (CLIP-domain serine proteases), SRPN (serine protease inhibitors), CASP (caspases), and IAP (inhibitors of apoptosis). The effector class includes SOD (superoxide dismutases), TEP (thioester-containing proteins), LYS (lysozymes), PPO (prophenoloxidases), PRDX (peroxidases), AMP (anti-microbial peptides), ML (MD2-like proteins), NIMROD (nimrod-related proteins), FREP (fibrinogen-related proteins), and CTL (C-type lectins).

Signatures of positive diversifying selection were searched for each category of immune genes as described in **Detection of positive selection signatures** section.

Evolutionary analysis of sex-determination genes

Protein sequences of *B. terrestris* genes including *feminizer* (*fem*), *feminizer* 1 (*fem* 1), and *transformer 2*, which are involved in the sex determination pathway, were used as queries to search against the newly sequenced genomes by locally installed BLAST (Gertz, et al. 2006) to identify their orthologs/paralogs in bumblebees. Before phylogenetic analysis, sequences were multiply aligned using MUSCLE (Edgar 2004). The evolutionary history of sex-determining genes in *Bombus* and related species was inferred using Maximum Likelihood with the JTT matrix-based model implemented in MEGA X (Jones, et al. 1992; Kumar, et al. 2018). The tree with the highest log likelihood (-6161.36) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 2.2)) with branch lengths measured in the number of amino acid substitutions per site. RELAX (Wertheim, et al. 2015) was employed to test whether the strength of natural selection was relaxed or intensified along a specified set of test branches. The spurious action of diversifying selection in a subset of branches was detected by aBSREL (Smith, et al. 2015). To further identify individual sites that were subject to episodic diversifying selection, the mixed effect model of evolution (MEME) was applied (Murrell, et al. 2012).

Identification of genes involved in the adaptation of bumblebees to high elevation

To identify genes involved in high-elevation adaptation, searches were conducted for genes undergoing positive selection in *B. superbus*, *B. waltoni*, *B. difficillimus*, and *B. skorikovi*, which were all collected at elevations > 4,000 m (Figure 1). First, universal single-copy orthologous groups were obtained, along with their respective multiple sequence alignments and Maximum Likelihood trees (described in the **Molecular evolution analysis on gene functional categories** section). Then, the improved branch-site model in the Codeml program of the PAML package was used to identify genes showing signatures of positive selection (Zhang, et al. 2005). In brief, *B. superbus*, *B. waltoni*, *B. difficillimus*, and *B. skorikovi* (all collected at elevations > 4,000 m) were assigned as the foreground branches and all the other bumblebee species (all collected at elevations < 2,000 m) as the background branches. A positive selection model that allowed a class of codons on the foreground branches to have $dN/dS > 1$ (model = 2, NSsites = 2, omega = 0.5|1.5, fix_omega = 0) was compared with a null model that constrained this class of sites to have $dN/dS = 1$ (model = 2, NSsites = 2, omega = 1, fix_omega = 1) using a likelihood ratio test and calculated a p-value for each comparison. Multiple comparisons were corrected for by using the Benjamini and Hochberg method and selected genes with an adjusted p-value < 0.05 as candidate positively selected genes (PSGs). Then, the Bayes Empirical Bayes (BEB) method (Yang, et al. 2005) was used to calculate posterior probabilities for site classes to identify codon positions that experienced positive selections ($dN/dS > 1$). Candidate PSGs that also contained codon positions showing significant BEB values (posterior probability >95%) were further analyzed using the software aBSREL (Smith, et al. 2015) to identify genes that show positive selection in at least two subgenera of high-elevation species but not in any of the low-elevation species. Such genes were believed to be PSGs involved in high-elevation adaptation. Finally, Codeml was used to estimate *dN*, *dS*, and *dN/dS* of these PSGs with the free ratio

model (model = 1, NSsites = 0). PSGs with *dS* >1, suggesting considerable saturation at the synonymous sites, were removed from downstream analysis to avoid false positives. Functional clues about the identified PSGs were obtained by BLAST searching against the UniProt database (https://www.uniprot.org) and by literature review.

Evolution of piRNA genes

Protein sequences for *Ago1*, *Armitage*, *Eggless*, *Gasz*, *Hen1*, *Maelstrom*, *Minotaur*, *Papi*, *Piwi/Aub*, *Qin*, *Shutdown*, *Spindle-E*, *Squash*, and *Trimmer* in *Apis mellifera* were downloaded from GenBank based on the dataset used by (Wang, et al. 2017). A BLAST protein database was built from the transcriptomes of each *Bombus* species and selected the top BLASTp hits for each species. We restricted our analyses to proteins that were present and had a single copy for all of the species.

Protein sequences were aligned using PSY-Coffee and automatically trimmed using G-Blocks while allowing for smaller final blocks and gap positions within the final blocks (Notredame, et al. 2000; Talavera and Castresana 2007). Phylogenies were estimated in MrBayes 3.2 with *Apis mellifera* set as the outgroup (Ronquist, et al. 2012). A mixed model for amino acid evolution was used. Each analysis ran for 10 million generations with the sampling frequency set to 1,000 with 3 heated chains, and 25% of the trees discarded as burnin.

The trimmed multiple alignments of single-copy orthologous groups containing piRNA genes, along with their phylogenies, were extracted from **Molecular evolution analysis on gene functional categories** section. Positive selection was detected by aBSREL (Smith, et al. 2015).

However, analysis of branch lengths and positive selection for 14 piRNA pathway genes across bumblebees found neither to be associated with genome size.

References

Alem S, Perry CJ, Zhu X, Loukola OJ, Ingraham T, Søvik E, Chittka L 2016. Associative mechanisms allow for social learning and cultural transmission of string pulling in an insect. PLoS Biol 14: e1002564.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic Local Alignment Search Tool. J Mol Biol 215: 403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389-3402.

Arbetman MP, Gleiser G, Morales CL, Williams P, Aizen MA 2017. Global decline of bumblebees is phylogenetically structured and inversely related to species range size and pathogen incidence. Proc Biol Sci 284: 20170204.

Ayala FJ, Coluzzi M 2005. Chromosome speciation: humans, Drosophila, and mosquitoes. Proc Natl Acad Sci U S A 102: 6535-6542. doi: 10.1073/pnas.0501847102

Ayasse M, Jarau S 2014. Chemical ecology of bumble bees. Annu Rev Entomol 59: 299-319.

1418 Bacon A, Harris A 2004. Hypoxia-inducible factors and hypoxic cell death in tumour
1419 physiology. Annals of medicine 36: 530-539.
1420 Baduel P, Quadrana L, Hunter B, Bomblies K, Colot V 2019. Relaxed purifying
1421 selection in autopolyploids drives transposable element over-accumulation which
1422 provides variants for local adaptation. Nat Commun 10: 1-10.
1423 Barbehenn RV 2001. Roles of peritrophic membranes in protecting herbivorous
1424 insects from ingested plant allelochemicals. Arch Insect Biochem Physiol 47: 86-99.
1425 Baron GL, Raine NE, Brown MJF 2017. General and species-specific impacts of a
1426 neonicotinoid insecticide on the ovary development and feeding of wild bumblebee
1427 queens. Proc Biol Sci 284: 20170123.
1428 Barribeau SM, Sadd BM, du Plessis L, Brown MJ, Buechel SD, Cappelle K, Carolan
1429 JC, Christiaens O, Colgan TJ, Erler S, Evans J, Helbing S, Karaus E, Lattorff HM,
1430 Marxer M, Meeus I, Napflin K, Niu J, Schmid-Hempel R, Smagghe G, Waterhouse
1431 RM, Yu N, Zdobnov EM, Schmid-Hempel P 2015. A depauperate immune repertoire
1432 precedes evolution of sociality in bees. Genome Biol 16: 83. doi: 10.1186/s13059-
1433 015-0628-y
1434 Bartomeus I, Ascher JS, Gibbs J, Danforth BN, Wagner DL, Hedtke SM, Winfree R
1435 2013. Historical changes in northeastern US bee pollinators related to shared
1436 ecological traits. Proc Natl Acad Sci U S A 110: 4656-4660.
1437 Behura SK, Severson DW 2012. Comparative analysis of codon usage bias and codon
1438 context patterns between dipteran and hymenopteran sequenced genomes. PLoS One
1439 7: e43111. doi: 10.1371/journal.pone.0043111
1440 Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J 2012. Hi–C: a
1441 comprehensive technique to capture the conformation of genomes. Methods 58: 268-
1442 276.
1443 Benson G 1999. Tandem repeats finder: a program to analyze DNA sequences.
1444 Nucleic acids research 27: 573-580.
1445 Biewer M, Schlesinger F, Hasselmann M 2015. The evolutionary dynamics of major
1446 regulators for sexual development among Hymenoptera species. Front Genet 6: 124.
1447 Bonetti B, Fu L, Moon J, Bedwell DM 1995. The Efficiency of Translation
1448 Termination is Determined by a Synergistic Interplay Between Upstream and
1449 Downstream Sequences inSaccharomyces cerevisiae. J Mol Biol 251: 0-345.
1450 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL
1451 2009. BLAST+: architecture and applications. BMC bioinformatics 10: 421.
1452 Cameron SA, Hines HM, Williams PH 2007. A comprehensive phylogeny of the
1453 bumble bees (Bombus). Biol J Linn Soc 91.
1454 Cameron SA, Lozier JD, Strange JP, Koch JB, Cordes N, Solter LF, Griswold TL
1455 2011. Patterns of widespread decline in North American bumble bees. Proc Natl Acad
1456 Sci U S A 108: 662-667.
1457 Cameron SA, Sadd BM 2019. Global Trends in Bumble Bee Health. Annu Rev
1458 Entomol 65.
1459 Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS,
1460 Yandell M 2008. MAKER: an easy-to-use annotation pipeline designed for emerging
1461 model organism genomes. Genome Res 18: 188-196.

1462 Capellagutierrez S, Sillamartinez JM, Gabaldon T 2009. trimAl: a tool for automated
1463 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972-
1464 1973.
1465 Casacuberta E, González J 2013. The impact of transposable elements in
1466 environmental adaptation. Mol Ecol 22: 1503-1517.
1467 Chen KC, Durand D, Farachcolton M 2000. NOTUNG: A Program for Dating Gene
1468 Duplications and Optimizing Gene Family Trees. J Comput Biol 7: 429-447.
1469 Cheong S-S, Hentschel L, Davidson AE, Gerrelli D, Davie R, Rizzo R, Pontikos N,
1470 Plagnol V, Moore AT, Sowden JC 2016. Mutations in CPAMD8 cause a unique form
1471 of autosomal-recessive anterior segment dysgenesis. Am J Hum Genet 99: 1338-
1472 1352.
1473 Criscuolo A, Gribaldo S 2010. BMGE (Block Mapping and Gathering with Entropy):
1474 a new software for selection of phylogenetic informative regions from multiple
1475 sequence alignments. BMC Evol Biol 10: 210-210.
1476 Csűros M 2008. Malin: maximum likelihood analysis of intron evolution in
1477 eukaryotes. Bioinformatics 24: 1538-1539.
1478 Da Lage J-L, Thomas GW, Bonneau M, Courtier-Orgogozo V 2019. Evolution of
1479 salivary glue genes in Drosophila species. BMC Evol Biol 19: 36.
1480 De Bie T, Cristianini N, Demuth JP, Hahn MW 2006. CAFE: a computational tool for
1481 the study of gene family evolution. Bioinformatics 22: 1269-1271.
1482 Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim
1483 MS, Machol I, Lander ES, Aiden AP 2017. De novo assembly of the Aedes aegypti
1484 genome using Hi-C yields chromosome-length scaffolds. Science 356: 92-95.
1485 Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS 2013. Ribosome profiling
1486 reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster.
1487 Elife 2: e01179.
1488 Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL
1489 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C
1490 Experiments. Cell systems 3: 95-98.
1491 Edgar RC 2004. MUSCLE: multiple sequence alignment with high accuracy and high
1492 throughput. Nucleic acids research 32: 1792-1797.
1493 Ellinghaus D, Kurtz S, Willhoeft U 2008. LTRharvest, an efficient and flexible
1494 software for de novo detection of LTR retrotransposons. BMC bioinformatics 9: 18-
1495 18.
1496 Enright, J. A 2002. An efficient algorithm for large-scale detection of protein
1497 families. Nucleic acids research 30: 1575-1584.
1498 Erlandson MA, Toprak U, Hegedus DD 2019. Role of the peritrophic matrix in
1499 insect-pathogen interactions. J Insect Physiol 117: 103894. doi:
1500 10.1016/j.jinsphys.2019.103894
1501 Evans JD, Aronstein KA, Chen Y, Hetru C, Imler J, Jiang H, Kanost MR, Thompson
1502 GJ, Zou Z, Hultmark D 2006. Immune pathways and defence mechanisms in honey
1503 bees Apis mellifera. Insect Mol Biol 15: 645-656.
1504 Falcon S, Gentleman R 2007. Using GOstats to test gene lists for GO term
1505 association. Bioinformatics 23: 257-258.

1506 Fontaine C, Dajoz I, Meriguet J, Loreau M 2005. Functional diversity of plant–
1507 pollinator interaction webs enhances the persistence of plant communities. PLoS Biol
1508 4: e1.
1509 Garibaldi LA, Steffan-Dewenter I, Winfree R, Aizen MA, Bommarco R, et al. 2013.
1510 Wild Pollinators Enhance Fruit Set of Crops Regardless of Honey Bee Abundance.
1511 Science 339: 1608-1611.
1512 Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF 2006. Composition-based
1513 statistics and translated nucleotide searches: Improving the TBLASTN module of
1514 BLAST. BMC Biology 4.
1515 Goulson D, Darvill B 2004. Niche overlap and diet breadth in bumblebees; are rare
1516 species more specialized in their choice of flowers? Apidologie 35: 55-63. doi:
1517 10.1051/apido:2003062
1518 Goulson D, Lye GC, Darvill B 2008. Decline and conservation of bumble bees. Annu
1519 Rev Entomol 53: 191-208. doi: 10.1146/annurev.ento.53.103106.093454
1520 Goulson D, Nicholls E, Botias C, Rotheray EL 2015. Bee declines driven by
1521 combined stress from parasites, pesticides, and lack of flowers. Science 347:
1522 1255957. doi: 10.1126/science.1255957
1523 Grixti JC, Wong LT, Cameron SA, Favret C 2009. Decline of bumble bees (Bombus)
1524 in the North American Midwest. Biol Conserv 142: 75-84. doi:
1525 10.1016/j.biocon.2008.09.027
1526 Gustafsson MV, Zheng X, Pereira T, Gradin K, Jin S, Lundkvist J, Ruas JL,
1527 Poellinger L, Lendahl U, Bondesson M 2005. Hypoxia requires notch signaling to
1528 maintain the undifferentiated cell state. Dev Cell 9: 617-628.
1529 Haas BJ, Delcher AL, Mount MSMS, Wortman JR, Smith RKW, Hannick LI, Maiti
1530 R, Ronning CM, Rusch DB, Town CD 2003. Improving the Arabidopsis genome
1531 annotation using maximal transcript alignment assemblies. Nucleic acids research 31:
1532 5654-5666.
1533 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger
1534 MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F,
1535 Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman
1536 N, Regev A 2013. De novo transcript sequence reconstruction from RNA-seq using
1537 the Trinity platform for reference generation and analysis. Nat Protoc 8.
1538 Han MV, Thomas GW, Lugo-Martinez J, Hahn MW 2013. Estimating gene gain and
1539 loss rates in the presence of error in genome assembly and annotation using CAFE 3.
1540 Mol Biol Evol 30: 1987-1997. doi: 10.1093/molbev/mst100
1541 Harris RS 2007. Improved Pairwise Alignment of Genomic DNA. PhD thesis
1542 Pennsylvania State Univ.
1543 Hasselmann M, Gempe T, Schiøtt M, Nunes-Silva CG, Otte M, Beye M 2008.
1544 Evidence for the evolutionary nascence of a novel sex determination pathway in
1545 honeybees. Nature 454: 519-522.
1546 Heavens D, Accinelli GG, Clavijo B, Clark MD 2015. A method to simultaneously
1547 construct up to 12 differently sized Illumina Nextera long mate pair libraries with
1548 reduced DNA input, time, and cost. Biotechniques 59: 42-45. doi:
1549 10.2144/000114310

1550 Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W 2004. Ten species in
1551 one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly
1552 Astraptes fulgerator. Proc Natl Acad Sci U S A 101: 14812-14817.
1553 Heng, Li, Durbin, Richard 2010. Fast and accurate long-read alignment with
1554 Burrows–Wheeler transform. Bioinformatics.
1555 Hines HM 2008. Historical biogeography, divergence times, and diversification
1556 patterns of bumble bees (Hymenoptera: Apidae: Bombus). Syst Biol 57: 58-75. doi:
1557 10.1080/10635150801898912
1558 Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS 2018. UFBoot2:
1559 Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35: 518-522.
1560 Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H
1561 2014. PASTEC: An Automatic Transposable Element Classification Tool. PLoS One
1562 9.
1563 Hofacker IL 2009. RNA Secondary Structure Analysis Using the Vienna RNA
1564 Package. Curr Protoc Bioinformatics 26.
1565 Jarvis ED, Mirarab S, Aberer AJ, Bo L, Houde P, et al. 2014. Whole-genome analyses
1566 resolve early branches in the tree of life of modern birds.
1567 Jones DT, Taylor WR, Thornton JM 1992. The rapid generation of mutation data
1568 matrices from protein sequences. Bioinformatics 8: 275-282.
1569 Jones PH, Binns D, Chang H, Fraser M, Li W, Mcanulla C, Mcwilliam H, Maslen J,
1570 Mitchell AL, Nuka G 2014. InterProScan 5: genome-scale protein function
1571 classification. Bioinformatics 30: 1236-1240.
1572 Jungreis I, Chan CS, Waterhouse RM, Fields G, Lin MF, Kellis M 2016. Evolutionary
1573 dynamics of abundant stop codon readthrough. Mol Biol Evol 33: 3108-3132.
1574 Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M
1575 2011. Evidence of abundant stop codon readthrough in Drosophila and other metazoa.
1576 Genome Res 21: 2096-2113.
1577 Junier T, Zdobnov EM 2010. The Newick utilities: high-throughput phylogenetic tree
1578 processing in the UNIX shell. Bioinformatics 26: 1669-1670. doi:
1579 10.1093/bioinformatics/btq243
1580 Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS 2017.
1581 ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods
1582 14: 587-589.
1583 Karpe SD, Jain R, Brockmann A, Sowdhamini R 2016. Identification of Complete
1584 Repertoire of Apis florea Odorant Receptors Reveals Complex Orthologous
1585 Relationships with Apis mellifera. Genome Biol Evol 8: 2879-2895.
1586 Katoh K, Misawa K, Kuma K, Miyata T 2002. MAFFT: a novel method for rapid
1587 multiple sequence alignment based on fast Fourier transform. Nucleic acids research
1588 30: 3059-3066.
1589 Kazutaka K, Standley DM 2013. MAFFT Multiple Sequence Alignment Software
1590 Version 7: Improvements in Performance and Usability. Mol Biol Evol 30: 772.
1591 Kerr JT, Pindar A, Galpern P, Packer L, Potts SG, Roberts SM, Rasmont P,
1592 Schweiger O, Colla SR, Richardson LL 2015. Climate change impacts on bumblebees
1593 converge across continents. Science 349: 177-180.

1594 Koch JB, Lozier J, Strange JP, Ikerd H, Griswold T, Cordes N, Solter L, Stewart I,
1595 Cameron SA 2015. USBombus, a database of contemporary survey data for North
1596 American Bumble Bees (Hymenoptera, Apidae, Bombus) distributed in the United
1597 States. Biodivers Data J: e6833. doi: 10.3897/BDJ.3.e6833
1598 Korf IF 2004. Gene finding in novel genomes. BMC bioinformatics 5: 59-59.
1599 Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A 2019. RAxML-NG: a fast,
1600 scalable and user-friendly tool for maximum likelihood phylogenetic inference.
1601 Bioinformatics 35: 4453-4455.
1602 Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simao FA, Pozdnyakov IA,
1603 Ioannidis P, Zdobnov EM 2015. OrthoDB v8: update of the hierarchical catalog of
1604 orthologs and the underlying free software. Nucleic acids research 43: D250-D256.
1605 Kubatko LS, Degnan JH 2007. Inconsistency of phylogenetic estimates from
1606 concatenated data under coalescence. Syst Biol 56: 17-24.
1607 Kumar S, Stecher G, Li M, Knyaz C, Tamura K 2018. MEGA X: Molecular
1608 Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35:
1609 1547-1549.
1610 Kunze J, Gumbert A 2001. The combined effect of color and odor on flower choice
1611 behavior of bumble bees in flower mimicry systems. Behavioral Ecology 12: 447-
1612 456.
1613 Lee TV, Sethi MK, Leonardi J, Rana NA, Buettner FF, Haltiwanger RS, Bakker H,
1614 Jafar-Nejad H 2013. Negative regulation of notch signaling by xylose. PLoS Genet 9.
1615 Leonhardt SD, Blüthgen N 2011. The same, but different: pollen foraging in
1616 honeybee and bumblebee colonies. Apidologie 43: 449-464. doi: 10.1007/s13592-
1617 011-0112-y
1618 Lhomme P, Hines HM 2019. Ecology and evolution of cuckoo bumble bees. Annals
1619 of the Entomological Society of America 112: 122-140.
1620 Li L, Stoeckert CJ, Roos DS 2003. OrthoMCL: Identification of Ortholog Groups for
1621 Eukaryotic Genomes. Genome Res 13: 2178-2189.
1622 Li W, Godzik A 2006. Cd-hit: a fast program for clustering and comparing large sets
1623 of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.
1624 Lin D, Hong P, Zhang S, Xu W, Jamal M, Yan K, Lei Y, Li L, Ruan Y, Fu ZF, Li G,
1625 Cao G 2018. Digestion-ligation-only Hi-C is an efficient and cost-effective method
1626 for chromosome conformation capture. Nat Genet 50: 754-763. doi: 10.1038/s41588-
1627 018-0111-2
1628 Lin MF, Jungreis I, Kellis M 2011. PhyloCSF: a comparative genomics method to
1629 distinguish protein coding and non-coding regions. Bioinformatics 27: i275-i282.
1630 Liu H, Jia Y, Sun X, Tian D, Hurst LD, Yang S 2017. Direct Determination of the
1631 Mutation Rate in the Bumblebee Reveals Evidence for Weak Recombination-
1632 Associated Mutation and an Approximate Rate Constancy in Insects. Mol Biol Evol
1633 34: 119-130. doi: 10.1093/molbev/msw226
1634 Liu J, Beqaj S, Yang Y, Honoré B, Schuger L 2001. Heterogeneous nuclear
1635 ribonucleoprotein-H plays a suppressive role in visceral myogenesis. Mechanisms of
1636 development 104: 79-87.

1637 Lomsadze A, Terhovhannisyan V, Chernoff YO, Borodovsky M 2005. Gene
1638 identification in novel eukaryotic genomes by self-training algorithm. Nucleic acids
1639 research 33: 6494-6506.
1640 Love RR, Weisenfeld NI, Jaffe DB, Besansky NJ, Neafsey DE 2016. Evaluation of
1641 DISCOVAR de novo using a mosquito sample for cost-effective short-read genome
1642 assembly. BMC Genomics 17. doi: 10.1186/s12864-016-2531-7
1643 Lowe TM, Eddy SR 1997. tRNAscan-SE: a program for improved detection of
1644 transfer RNA genes in genomic sequence. Nucleic acids research 25: 955-964.
1645 Maddison WP 1997. Gene trees in species trees. Syst Biol 46: 523-536.
1646 Manjila SB, Kuruvilla M, Ferveur J-F, Sane SP, Hasan G 2019. Extended flight bouts
1647 require disinhibition from GABAergic mushroom body neurons. Curr Biol 29: 283-
1648 293. e285.
1649 Martin CD, Fountain MT, Brown MJF 2019. Varietal and seasonal differences in the
1650 effects of commercial bumblebees on fruit quality in strawberry crops. Agric Ecosyst
1651 Environ 281: 124-133. doi: 10.1016/j.agee.2019.04.007
1652 Mendes FK, Hahn MW 2016. Gene tree discordance causes apparent substitution rate
1653 variation. Syst Biol 65: 711-721.
1654 Mendes FK, Hahn MW 2018. Why concatenation fails near the anomaly zone. Syst
1655 Biol 67: 158-169.
1656 Merzendorfer H, Zimoch L 2003. Chitin metabolism in insects: structure, function
1657 and regulation of chitin synthases and chitinases. J Exp Biol 206: 4393-4412. doi:
1658 10.1242/jeb.00709
1659 Meyer-Rochow V 2019. Eyes and Vision of the Bumblebee: a Brief Review on how
1660 Bumblebees Detect and Perceive Flowers. Journal of Apiculture 34: 107-115.
1661 Minh BQ, Hahn M, Lanfear R 2018. New methods to calculate concordance factors
1662 for phylogenomic datasets. BioRXiv: 487801.
1663 Minh BQ, Hahn MW, Lanfear R 2020a. New methods to calculate concordance
1664 factors for phylogenomic datasets. Mol Biol Evol. doi: 10.1093/molbev/msaa106
1665 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler
1666 A, Lanfear R 2020b. IQ-TREE 2: New models and efficient methods for phylogenetic
1667 inference in the genomic era. Mol Biol Evol 37: 1530-1534.
1668 Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK 2012.
1669 Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet 8.
1670 Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE,
1671 Amon J, Arca B, Arensburger P, Artemov G, Assour LA, Basseri H, Berlin A, Birren
1672 BW, Blandin SA, Brockman AI, Burkot TR, Burt A, Chan CS, Chauve C, Chiu JC,
1673 Christensen M, Costantini C, Davidson VL, Deligianni E, Dottorini T, Dritsou V,
1674 Gabriel SB, Guelbeogo WM, Hall AB, Han MV, Hlaing T, Hughes DS, Jenkins AM,
1675 Jiang X, Jungreis I, Kakani EG, Kamali M, Kemppainen P, Kennedy RC,
1676 Kirmitzoglou IK, Koekemoer LL, Laban N, Langridge N, Lawniczak MK, Lirakis M,
1677 Lobo NF, Lowy E, MacCallum RM, Mao C, Maslen G, Mbogo C, McCarthy J,
1678 Michel K, Mitchell SN, Moore W, Murphy KA, Naumenko AN, Nolan T, Novoa EM,
1679 O'Loughlin S, Oringanje C, Oshaghi MA, Pakpour N, Papathanos PA, Peery AN,
1680 Povelones M, Prakash A, Price DP, Rajaraman A, Reimer LJ, Rinker DC, Rokas A,

Russell TL, Sagnon N, Sharakhova MV, Shea T, Simao FA, Simard F, Slotman MA, Somboon P, Stegniy V, Struchiner CJ, Thomas GW, Tojo M, Topalis P, Tubio JM, Unger MF, Vontas J, Walton C, Wilding CS, Willis JH, Wu YC, Yan G, Zdobnov EM, Zhou X, Catteruccia F, Christophides GK, Collins FH, Cornman RS, Crisanti A, Donnelly MJ, Emrich SJ, Fontaine MC, Gelbart W, Hahn MW, Hansen IA, Howell PI, Kafatos FC, Kellis M, Lawson D, Louis C, Luckhart S, Muskavitch MA, Ribeiro JM, Riehle MA, Sharakhov IV, Tu Z, Zwiebel LJ, Besansky NJ 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science 347: 1258522. doi: 10.1126/science.1258522

Notredame C, Higgins DG, Heringa J 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205-217.

Oppenheim S, Cao X, Rueppel O, Krongdang S, Phokasem P, DeSalle R, Goodwin S, Xing J, Chantawannakul P, Rosenfeld JA 2020. Whole Genome Sequencing and Assembly of the Asian Honey Bee Apis dorsata. Genome Biol Evol 12: 3677-3683.

Owen, Robin E 1983. Chromosome numbers of 15 North American bumble bee species (Hymenoptera, Apidae, Bombini). Canadian Journal of Genetics and Cytology 25: 26-29.

Owen RE, Richards KW, Wilkes A 1995. Chromosome Numbers and Karyotypic Variation in Bumble Bees (Hymenoptera: Apidae; Bombini). J Kansas Entomol Soc 68.

Pan T 2013. Adaptive translation as a mechanism of stress response and adaptation. Annu Rev Genet 47: 121-137. doi: 10.1146/annurev-genet-111212-133522

Park D, Jung JW, Choi B-S, Jayakodi M, Lee J, Lim J, Yu Y, Choi Y-S, Lee M-L, Park Y, Choi I-Y, Yang T-J, Edwards OR, Nah G, Kwon HW 2015. Uncovering the novel characteristics of Asian honey bee, Apis cerana , by whole genome sequencing. BMC Genomics 16.

Pashalidou FG, Lambert H, Peybernes T, Mescher MC, De Moraes CM 2020. Bumble bees damage plant leaves and accelerate flower production when pollen is scarce. Science 368: 881-884.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino DR, Haussler D 2011. Cactus: Algorithms for genome multiple sequence alignment. Genome Res 21: 1512-1528.

Pease JB, Haak DC, Hahn MW, Moyle LC 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol 14.

Persson AS, Rundlöf M, Clough Y, Smith HG 2015. Bumble bees show trait-dependent vulnerability to landscape simplification. Biodivers Conserv 24: 3469-3489. doi: 10.1007/s10531-015-1008-3

Pertea M, Kim D, Pertea G, Leek JT, Salzberg SL 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11: 1650-1667.

Pigna E, Simonazzi E, Sanna K, Bernadzki KM, Proszynski T, Heil C, Palacios D, Adamo S, Moresi V 2019. Histone deacetylase 4 protects from denervation and skeletal muscle atrophy in a murine model of amyotrophic lateral sclerosis. EBioMedicine 40: 717-732.

1724 Potts SG, Biesmeijer JC, Kremen C, Neumann P, Schweiger O, Kunin WE 2010.
1725 Global pollinator declines: trends, impacts and drivers. Trends Ecol Evol 25: 345-353.
1726 doi: 10.1016/j.tree.2010.01.007
1727 Price AL, Jones NC, Pevzner PA 2005. De novo identification of repeat families in
1728 large genomes. Bioinformatics 21: 351-358.
1729 Prufer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J 2008. PatMaN:
1730 Rapid alignment of short sequences to large databases. Bioinformatics 24: 1530-1531.
1731 Rajput B, Pruitt KD, Murphy TD 2019. RefSeq curation and annotation of stop codon
1732 recoding in vertebrates. Nucleic acids research 47: 594-606.
1733 Rho M, Tang H 2009. MGEScan-non-LTR: computational identification and
1734 classification of autonomous non-LTR retrotransposons in eukaryotic genomes.
1735 Nucleic acids research 37.
1736 Robertson HM, Wanner KW 2006. The chemoreceptor superfamily in the honey bee,
1737 Apis mellifera: Expansion of the odorant, but not gustatory, receptor family. Genome
1738 Res 16: 1395-1403.
1739 Ronquist F, Teslenko M, Mark Pvd, Ayres DL, Darling A, Höhna S, Larget B, Liu L,
1740 Suchard MA, Huelsenbeck JP 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic
1741 Inference and Model Choice Across a Large Model Space. Syst Biol 61.
1742 Ruedenauer FA, Spaethe J, Leonhardt SD 2015. How to know which food is good for
1743 you: bumblebees use taste to discriminate between different concentrations of food
1744 differing in nutrient content. J Exp Biol 218: 2233-2240. doi: 10.1242/jeb.118554
1745 Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG 2007.
1746 Dynamic evolution of the innate immune system in Drosophila. Nat Genet 39: 1461-
1747 1468. doi: 10.1038/ng.2007.60
1748 Sadd BM, Barribeau SM, Bloch G, Graaf DCd, Dearden P, Elsik CG, Gadau J,
1749 Grimmelikhuijzen CJ, Hasselmann M, Lozier JD, Robertson HM, Smagghe G, Stolle
1750 E, Vaerenbergh MV, Waterhouse RM, Bornberg-Bauer E, Klasberg S, Bennett AK,
1751 Câmara F, Guigó R, Hoff K, Mariotti M, Munoz-Torres M, Murphy T, Santesmasses
1752 D, Amdam GV, Beckers M, Beye M, Biewer M, M M 2015. The genomes of two key
1753 bumblebee species with primitive eusocial organization. Genome Biol 16.
1754 Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L 2014. BESST-efficient
1755 scaffolding of large fragmented assemblies. BMC bioinformatics 15: 281.
1756 Sanderson MJ 2003. r8s: inferring absolute rates of molecular evolution and
1757 divergence times in the absence of a molecular clock. Bioinformatics 19: 301-302.
1758 Schrader L, Schmitz J 2019. The impact of transposable elements in adaptive
1759 evolution. Mol Ecol 28: 1537-1549.
1760 Shirato K, Nakajima K, Korekane H, Takamatsu S, Gao C, Angata T, Ohtsubo K,
1761 Taniguchi N 2010. Hypoxic regulation of glycosylation via the N-acetylglucosamine
1762 cycle. Journal of clinical biochemistry and nutrition 48: 20-25.
1763 Sikora A, Kelm M 2012. Flower Preferences of the Wrocław Botanical Garden
1764 Bumblebees (Bombus spp.). Journal of Apicultural Science 56: 27-36. doi:
1765 10.2478/v10289-012-0021-y
1766 Slater G, Birney E 2005. Automated generation of heuristics for biological sequence
1767 comparison. BMC bioinformatics 6: 31-31.

1768 Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Pond SLK 2015. Less Is
1769 More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of
1770 Episodic Diversifying Selection. Mol Biol Evol 32: 1342-1353.
1771 Somme L, Vanderplanck M, Michez D, Lombaerde I, Moerman R, Wathelet B,
1772 Wattiez R, Lognay G, Jacquemart AL 2015. Pollen and nectar quality drive the major
1773 and minor floral choices of bumble bees. Apidologie 46: 92-106.
1774 Soroye P, Newbold T, Kerr J 2020. Climate change contributes to widespread
1775 declines among bumble bees across continents. Science 367: 685-688.
1776 Stamatakis A 2014. RAxML version 8: a tool for phylogenetic analysis and post-
1777 analysis of large phylogenies. Bioinformatics 30: 1312-1313.
1778 Stanke M, Schoffmann O, Morgenstern B, Waack S 2006. Gene prediction in
1779 eukaryotes with a generalized hidden Markov model that uses hints from external
1780 sources. BMC bioinformatics 7: 62-62.
1781 Steinbiss S, Willhoeft U, Gremme G, Kurtz S 2009. Fine-grained annotation and
1782 classification of de novo predicted LTR retrotransposons. Nucleic acids research 37:
1783 7002-7013.
1784 Subramanian S 2008. Nearly Neutrality and the Evolution of Codon Usage Bias in
1785 Eukaryotic Genomes. Genetics 178: 2429-2432.
1786 Suyama M, Torrents D, Bork P 2006. PAL2NAL: robust conversion of protein
1787 sequence alignments into the corresponding codon alignments. Nucleic acids research
1788 34: 609-612.
1789 Talavera G, Castresana J 2007. Improvement of Phylogenies after Removing
1790 Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst
1791 Biol 56: 564-577.
1792 Tesler G 2002. Efficient algorithms for multichromosomal genome rearrangements. J
1793 Comp Sys Sci 65: 587-609.
1794 Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K,
1795 Anstead CA, Ayoub NA, Batterham P, Bellair M, Binford GJ, Chao H, Chen YH,
1796 Childers C, Dinh H, Doddapaneni HV, Duan JJ, Dugan S, Esposito LA, Friedrich M,
1797 Garb J, Gasser RB, Goodisman MAD, Gundersen-Rindal DE, Han Y, Handler AM,
1798 Hatakeyama M, Hering L, Hunter WB, Ioannidis P, Jayaseelan JC, Kalra D, Khila A,
1799 Korhonen PK, Lee CE, Lee SL, Li Y, Lindsey ARI, Mayer G, McGregor AP,
1800 McKenna DD, Misof B, Munidasa M, Munoz-Torres M, Muzny DM, Niehuis O,
1801 Osuji-Lacy N, Palli SR, Panfilio KA, Pechmann M, Perry T, Peters RS, Poynton HC,
1802 Prpic NM, Qu J, Rotenberg D, Schal C, Schoville SD, Scully ED, Skinner E, Sloan
1803 DB, Stouthamer R, Strand MR, Szucsich NU, Wijeratne A, Young ND, Zattara EE,
1804 Benoit JB, Zdobnov EM, Pfrender ME, Hackett KJ, Werren JH, Worley KC, Gibbs
1805 RA, Chipman AD, Waterhouse RM, Bornberg-Bauer E, Hahn MW, Richards S 2020.
1806 Gene content evolution in the arthropods. Genome Biol 21: 15. doi: 10.1186/s13059-
1807 019-1925-7
1808 Véga C, R. Sachleben Jr L, Gozal D, Gozal E 2006. Differential metabolic adaptation
1809 to acute and long-term hypoxia in rat primary cortical astrocytes. J Neurochem 97:
1810 872-883.

1811 Velthuis HHW, Van Doorn A 2006. A century of advances in bumblebee
1812 domestication and the economic and environmental aspects of its commercialization
1813 for pollination. Apidologie 37: 421-451. doi: 10.1051/apido:2006019
1814 Vicario S, Moriyama EN, Powell JR 2007. Codon usage in twelve species of
1815 Drosophila. BMC Evol Biol 7: 226. doi: 10.1186/1471-2148-7-226
1816 von Grotthuss M, Ashburner M, Ranz JM 2010. Fragile regions and not functional
1817 constraints predominate in shaping gene organization in the genus Drosophila.
1818 Genome Res 20: 1084-1096.
1819 Walski T, De Schutter K, Van Damme EJM, Smagghe G 2017. Diversity and
1820 functions of protein glycosylation in insects. Insect Biochem Mol Biol 83: 21-34. doi:
1821 10.1016/j.ibmb.2017.02.005
1822 Wang G, Yin H, Li B, Yu C, Wang F, Xu X, Cao J, Bao Y, Wang L, Abbasi AA
1823 2019. Characterization and identification of long non-coding RNAs based on feature
1824 relationship. Bioinformatics 35: 2949-2956.
1825 Wang L, Park HJ, Dasari S, Wang S, Kocher JA, Li W 2013. CPAT: Coding-
1826 Potential Assessment Tool using an alignment-free logistic regression model. Nucleic
1827 acids research 41.
1828 Wang W, Ashby R, Ying H, Maleszka R, Foret S 2017. Contrasting sex-and caste-
1829 dependent piRNA profiles in the transposon depleted haplodiploid honeybee Apis
1830 mellifera. Genome Biol Evol 9: 1341-1356.
1831 Wang Y, Tang H, D DJ, Xu T, Li J, Wang X, Tae-ho L, Jin H, Barry M, Hui G 2012.
1832 MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and
1833 collinearity. Nucleic acids research 40: e49.
1834 Waterhouse RM, Lazzaro BP, Sackton TB 2020. Characterization of Insect Immune
1835 Systems from Genomic Data. Immunity in Insects.
1836 Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G,
1837 Kriventseva EV, Zdobnov EM 2018. BUSCO applications from quality assessments
1838 to gene prediction and phylogenomics. Mol Biol Evol 35: 543-548.
1839 Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Weinstock GM, Robinson
1840 GE, Worley KC, Evans JD, Maleszka R, Robertson HM 2006. Insights into social
1841 insects from the genome of the honeybee Apis mellifera. Nature 443.
1842 Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K 2015. RELAX:
1843 Detecting Relaxed Selection in a Phylogenetic Framework. Mol Biol Evol 32: 820-
1844 832.
1845 Wilfert L, Gadau J, Schmid-Hempel P 2007. Variation in genomic recombination
1846 rates among animal taxa and the case of social insects. Heredity (Edinb) 98: 189-197.
1847 doi: 10.1038/sj.hdy.6800950
1848 Williams, Paul H 1985. A preliminary cladistic investigation of relationships among
1849 the bumble bees (Hymenoptera, Apidae). Systematic Entomology 10: 239-255.
1850 Williams DS, Lopes VS 2011. The many different cellular functions of MYO7A in
1851 the retina. Biochem Soc Trans 39: 1207-1210.
1852 Williams P, Colla S, Xie Z 2009. Bumblebee vulnerability: common correlates of
1853 winners and losers across three continents. Conserv Biol 23: 931-940. doi:
1854 10.1111/j.1523-1739.2009.01176.x

Williams PH 1998. An Annotated Checklist of Bumble Bees with an Analysis of Patterns of Description (Hymenoptera: Apidae, Bombini). Bulletin of the Natural History Museum Entomology 67: 79-152.

Williams PH 1994. Phylogenetic relationships among bumble bees (Bombus Latr.): a reappraisal of morphological evidence. Systematic Entomology 19: 327-344.

Williams PH, Berezin MV, Cannings SG, Cederberg B, Ødegaard F, Rasmussen C, Richardson LL, Rykken J, Sheffield CS, Thanoosing C 2019. The arctic and alpine bumblebees of the subgenus Alpinobombus revised from integrative assessment of species' gene coalescents and morphology (Hymenoptera, Apidae, Bombus). Zootaxa 4625: 1-68.

Williams PH, Cameron SA, Hines HM, Cederberg B, Rasmont P 2008. A simplified subgeneric classification of the bumblebees (genusBombus). Apidologie 39: 46-74. doi: 10.1051/apido:2007052

Williams PH, Huang J, Rasmont P, An J 2016. Early-diverging bumblebees from across the roof of the world: the high-mountain subgenus Mendacibombus revised from species' gene coalescents and morphology (Hymenoptera, Apidae). Zootaxa 4204: 1-72. doi: 10.11646/zootaxa.4204.1.1

Williams PH, Lobo JM, Meseguer AS 2018. Bumblebees take the high road: climatically integrative biogeography shows that escape from Tibet, not Tibetan uplift, is associated with divergences of present-day Mendacibombus. Ecography 41: 461-477. doi: 10.1111/ecog.03074

Williams PH, Osborne JL 2009. Bumblebee vulnerability and conservation world-wide. Apidologie 40: 367-387. doi: 10.1051/apido/2009025

Yang Z 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24: 1586-1591.

Yang Z, Wong WSW, Nielsen R 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol Biol Evol 22: 1107-1118.

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L 2006. WEGO: a web tool for plotting GO annotations. Nucleic acids research 34: 293-297.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8: 28-36.

Zhang J, Nielsen R, Yang Z 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. Mol Biol Evol 22: 2472-2479.

Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J, Wai CM, Zheng C, Shi Y, Chen S, Xu X, Yue J, Nelson DR, Huang L, Li Z, Xu H, Zhou D, Wang Y, Hu W, Lin J, Deng Y, Pandey N, Mancini M, Zerpa D, Nguyen JK, Wang L, Yu L, Xin Y, Ge L, Arro J, Han JO, Chakrabarty S, Pushko M, Zhang W, Ma Y, Ma P, Lv M, Chen F, Zheng G, Xu J, Yang Z, Deng F, Chen X, Liao Z, Zhang X, Lin Z, Lin H, Yan H, Kuang Z, Zhong W, Liang P, Wang G, Yuan Y, Shi J, Hou J, Lin J, Jin J, Cao P, Shen Q, Jiang Q, Zhou P, Ma Y, Zhang X, Xu R, Liu J, Zhou Y, Jia H, Ma Q, Qi R, Zhang Z, Fang J, Fang H, Song J, Wang M, Dong G, Wang G,

1899 Chen Z, Ma T, Liu H, Dhungana SR, Huss SE, Yang X, Sharma A, Trujillo JH,
1900 Martinez MC, Hudson M, Riascos JJ, Schuler M, Chen LQ, Braun DM, Li L, Yu Q,
1901 Wang J, Wang K, Schatz MC, Heckerman D, Van Sluys MA, Souza GM, Moore PH,
1902 Sankoff D, VanBuren R, Paterson AH, Nagai C, Ming R 2018. Allele-defined
1903 genome of the autopolyploid sugarcane Saccharum spontaneum L. Nat Genet 50:
1904 1565-1573. doi: 10.1038/s41588-018-0237-2
1905 Zhao X, Xu W, Schaack S, Sun C 2019. Genome-wide identification of accessible
1906 chromatin regions in bumblebee (Bombus terrestris) by ATAC-seq. BioRXiv:
1907 818211.
1908 Zhou X, Rokas A, Berger SL, Liebig J, Ray A, Zwiebel LJ 2015. Chemoreceptor
1909 Evolution in Hymenoptera and Its Implications for the Evolution of Eusociality.
1910 Genome Biol Evol 7: 2407-2416.
1911 Zhou X, Slone J, Rokas A, Berger SL, Liebig J, Ray A, Reinberg D, Zwiebel LJ
1912 2012. Phylogenetic and Transcriptomic Analysis of Chemosensory Receptors in a
1913 Pair of Divergent Ant Species Reveals Sex-Specific Signatures of Odor Coding. PLoS
1914 Genet 8.
1915 Zhu KY, Merzendorfer H, Zhang W, Zhang J, Muthukrishnan S 2016. Biosynthesis,
1916 turnover, and functions of chitin in insects. Annu Rev Entomol 61: 177-196.
1917