Article

Current Biology

A Single Gene Causes Thelytokous Parthenogenesis, the Defining Feature of the Cape Honeybee *Apis mellifera capensis*

Highlights

- Gene GB45239 causes thelytoky (virgin birth) in *A. m. capensis* honeybees
- Variants in GB45239 consistently co-segregate with thelytoky across populations
- GB45239 is downregulated in the ovaries of thelytokous bees
- GB45239 likely affects chromosome segregation, which results in a faulty meiosis

Authors

Boris Yagound, Kathleen A. Dogantzis, Amro Zayed, ..., Orly Dim, Gabriele Buchmann, Benjamin P. Oldroyd

Correspondence

boris.yagound@sydney.edu.au (B.Y.), benjamin.oldroyd@sydney.edu.au (B.P.O.)

In Brief

Reversions from sexual to asexual reproduction are rare. Yagound et al. identify a genetic variant that is consistently associated with asexual reproduction in the honeybee at individual and population scales. The gene likely orchestrates the fusion of maternal pronuclei in the egg to restore diploidy without need of a sperm.

Yagound et al., 2020, Current Biology *30*, 1–12
June 22, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cub.2020.04.033

CellPress

Please cite this article in press as: Yagound et al., A Single Gene Causes Thelytokous Parthenogenesis, the Defining Feature of the Cape Honeybee *Apis mellifera capensis*, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.04.033

Current Biology

CellPress

Article

A Single Gene Causes Thelytokous Parthenogenesis, the Defining Feature of the Cape Honeybee *Apis mellifera capensis*

Boris Yagound,[1,5,*] Kathleen A. Dogantzis,[2] Amro Zayed,[2] Julianne Lim,[1] Paul Broekhuyse,[1] Emily J. Remnant,[1] Madeleine Beekman,[1] Michael H. Allsopp,[3] Sarah E. Aamidor,[1] Orly Dim,[4] Gabriele Buchmann,[1] and Benjamin P. Oldroyd[1,*]

[1]Behaviour, Ecology and Evolution Laboratory, School of Life and Environmental Sciences, Science Road, University of Sydney, Sydney, NSW 2006, Australia

[2]Department of Biology, Faculty of Science, 4700 Keele Street, York University, Toronto, ON M3J 1P3, Canada

[3]Honeybee Research Section, ARC-Plant Protection Research Institute, Private Bag X5017, Stellenbosch 7600, South Africa

[4]Structural Proteomics Unit, Faculty of Biochemistry, 234 Herzl Street, Weizmann Institute of Science, Rehovot 7610001, Israel

[5]Lead Contact

*Correspondence: boris.yagound@sydney.edu.au (B.Y.), benjamin.oldroyd@sydney.edu.au (B.P.O.)

https://doi.org/10.1016/j.cub.2020.04.033

SUMMARY

In honeybees, the ability of workers to produce daughters asexually, i.e., thelytokous parthenogenesis, is restricted to a single subspecies inhabiting the Cape region of South Africa, *Apis mellifera capensis*. Thelytoky has unleashed new selective pressures and the evolution of traits such as social parasitism, invasiveness, and social cancer. Thelytoky arises from an abnormal meiosis that results in the fusion of two maternal pronuclei, restoring diploidy in newly laid eggs. The genetic basis underlying thelytoky is disputed. To resolve this controversy, we generated a backcross between thelytokous *A. m. capensis* and non-thelytokous *A. m. scutellata* from the neighboring population and looked for evidence of genetic markers that co-segregated with thelytokous reproduction in 49 backcross females. We found that markers associated with the gene GB45239 on chromosome 11, including non-synonymous variants, showed consistent co-segregation with thelytoky, whereas no other region did so. Alleles associated with thelytoky were present in all *A. m. capensis* genomes examined but were absent from all other honeybees worldwide including *A. m. scutellata*. GB45239 is derived in *A. m. capensis* and has a putative role in chromosome segregation. It is expressed in ovaries and is downregulated in thelytokous bees, likely because of polymorphisms in the promoter region. Our study reveals how mutations affecting the sequence and/or expression of a single gene can change the reproductive mode of a population.

INTRODUCTION

Sexual reproduction is ubiquitous among eukaryotes, yet switches to asexual reproduction are frequent. Transitions from sex to asex can involve different mechanisms, including endosymbionts, hybridization events, and genetic factors [1, 2]. Changes in reproductive mode are of great evolutionary significance, and much attention has been given to evaluating the relative benefits of sex and asex [3–6]. By contrast, the molecular mechanisms underpinning the transition between alternative modes of reproduction are largely unknown.

In haplo-diploid species such as honeybees it is usual for unfertilized eggs to develop as haploid males asexually by arrhenotokous parthenogenesis, whereas diploid females always arise from sexual reproduction [7, 8]. *Apis mellifera capensis* (hereafter Capensis), a honeybee subspecies that is confined to the southern two provinces of South Africa, is a remarkable exception [9, 10]. The major defining feature of Capensis is its highly unusual mode of asexual reproduction: thelytokous parthenogenesis. In Capensis, unfertilized eggs laid by workers develop as diploid females [11]. Two non-sister cells of the four products of an otherwise normal meiosis fuse within the egg to restore diploidy, as if one nucleus acted as a sperm producing a daughter [12–14]. Thelytoky in Capensis has profound consequences, because workers have the possibility of being genetically reincarnated as a queen [15–17] and regularly parasitize their own or an unrelated colony [18–21] with their clonal eggs [22].

Capensis is distinct from *A. m. scutellata* (hereafter Scutellata), a population that has a broad range across northern provinces of South Africa and in countries further north [10]. The two subspecies are separated by a hybrid zone, and colonies of both subspecies do not persist within each other's range [22–24]. Nevertheless, the two subspecies are very similar genetically (genome-wide F_{ST} ~0.05) [25]. Likely as a consequence of the evolutionary opportunities made possible by worker thelytoky [26], Capensis has evolved a complex of behaviors that predispose them to social parasitism [27–29]. The ultimate manifestation of Capensis social parasitism is a clonal lineage of workers that infests the commercial Scutellata population of South Africa [30].

There has been a long history of attempts to understand the genetic basis of thelytoky in Capensis. Crosses between thelytokous Capensis and arrhenotokous honeybees produced worker progeny that were either thelytokous or arrhenotokous but never both [10, 31]. These results were interpreted as a sign of qualitative inheritance, although they could potentially be explained under a polygenic model. Subsequently, a backcross between Capensis and arrhenotokous honeybees, suggested that a single *thelytoky* (*th*) locus located on chromosome 13 controls thelytoky [32]. A subsequent backcross experiment mapped a presumptive thelytoky-causing gene to the same region, GB48238 (*Transcription factor CP2-like protein 1*) [33]. However, population surveys have now made it clear that this gene is not the switch between arrhenotoky and thelytoky [25, 34, 35].

A genome-wide association study within a single Capensis colony headed by a naturally mated queen identified a second candidate *th* locus on chromosome 1, GB46427 [35, 36]. GB46427 is linked to *Ecdysis-triggering hormone receptor* (*Ethr*) within a non-recombining region of the genome [36]. A non-synonymous single-nucleotide polymorphism (SNP) within GB46427 was associated with the mode of parthenogenesis in the studied colony. However, the association between mode of reproduction and genotype at the putative thelytoky locus did not hold after more extensive sampling of the Capensis and Scutellata populations [37]. It remains possible that GB46427 plays a role in other traits specific to Capensis, e.g., their body coloration, though the proposed over-dominant balanced polymorphism [36] controlling the phenotype makes this unlikely.

Finally, a population study sought to identify genomic regions that showed strong genetic divergence between the thelytokous Capensis population and other, arrhenotokous, honeybee populations [25]. There was generally low genetic divergence of SNPs between Capensis and other African subspecies (genome-wide $F_{ST} = 0.051–0.056$). However, 12 regions on eight chromosomes showed extreme divergence ($F_{ST} > 0.8$) and signatures of selection in Capensis. These 12 regions are candidates for loci that influence thelytoky, though it is clear that, if thelytoky is indeed controlled by a single locus, 11 of the 12 regions are likely associated not with thelytoky per se but with other Capensis-specific traits [25]. Of particular interest were two 0.5 Mb blocks on chromosome 1, one of them intersecting with *Ethr* and GB46427 identified in [36], and a 1 Mb block on chromosome 11 that contained an uncharacterized protein-coding gene LOC100576557 (GB45239) and that showed the strongest divergence between Capensis and Scutellata.

We assumed that the underlying genetic basis of thelytoky in Capensis are loci that must be homozygous for a recessive thelytoky-causing allele (*th*) to produce the thelytoky phenotype. We based this assumption on the observation that F_1 worker progeny of crosses between Capensis and arrhenotokous subspecies show negligible evidence of thelytoky, whereas backcrosses to the putative recessive parent (Capensis) generate both thelytokous and arrhenotokous workers [31–34]. The phenotype could involve one or several loci [34], that could be detected by a backcross experiment. We used artificial insemination to generate an F_1 cross between thelytokous Capensis and arrhenotokous Scutellata, which we backcrossed to a Capensis male (Figure 1A). We then used whole-genome sequencing of the grandparents, parents, and 49 of the backcross workers' female offspring (which by definition were thelytokously produced) to identify a single locus that was consistently associated with the presence of thelytoky.

RESULTS

A Region on Chromosome 11 Is Strongly Associated with Thelytoky

We sequenced 53 genomes with 47.47 ± 0.74-fold (mean ± SE, range 36.20–59.28) coverage (Table S1) from our backcross. We then identified a set of 62,526 SNPs that satisfied the following criteria: (1) homozygous in the 49 backcross workers' female offspring, (2) heterozygous in the workers' F_1 hybrid mother, (3) present in the workers' backcross Capensis father, (4) homozygous in the workers' thelytokous Capensis grandmother, and (5) absent in the workers' Scutellata grandfather. There was an average of 3,907.88 ± 363.57 SNPs per chromosome (Table S2). We used a sliding window approach to identify putative regions associated with thelytoky and calculated a frequency score for each SNP, defined as the proportion of workers' offspring that were homozygous for the allele found in the Capensis father out of the total number of workers' offspring. The average number of 100 SNPs sliding windows per chromosome was 155.81 ± 14.50 (total 2,493) (Table S2). For SNPs that are unlinked to *th*, we would expect frequencies of 0.5 (Figure S1). As expected, the average SNP frequency score was 0.512 ± 0.001 across the genome (Figure 1B). Likewise, the average frequency score per window was 0.512 ± 0.002.

Of the 62,526 informative SNPs, 549 (0.88%) were high frequency (HF) (i.e., had a frequency score ≥ 0.95) in the progeny of backcross workers (Table S2). All but two of the HF SNPs were located in a 1.5 Mb region on chromosome 11 (Figure 1B; Table S2). Chromosome 11 had an otherwise unremarkable number of SNPs (4,054) and windows (162) compared with other chromosomes (Table S2). A 105 kb subset of this region, located between positions 14,530,850 and 14,635,532, encompassed 254 SNPs, of which 244 were HF (44.44% of all HF SNPs) (Figure S2). The top three windows, located between positions 14,586,727 and 14,610,124, had 98–100 HF SNPs (out of a possible maximum of 100) and an average frequency score across all 100 SNPs of 0.977–0.978.

To further investigate how these HF SNPs deviated from frequencies expected for SNPs unlinked to *th*, we conducted a binomial test for each informative SNP, to determine whether the observed frequencies were significantly greater than 0.5. All of the 549 previously identified HF-SNPs fell above the 99% percentile genome-wide of Benjamini-Hochberg-corrected p values and were deemed highly significant (Figure 1C; Table S2). Our candidate region on chromosome 11 had 24 overlapping windows with more than 50 highly significant HF SNPs, of which the top windows had 98–100 highly significant HF SNPs (Table S2). The only two other regions with a window containing a HF SNP were located, respectively, on chromosomes 5 and 9. However, both windows had only one highly significant HF SNP (average frequency score across all 100 SNPs of 0.542 and 0.511, respectively; Table S2). It is therefore highly unlikely that these regions are associated with thelytoky.

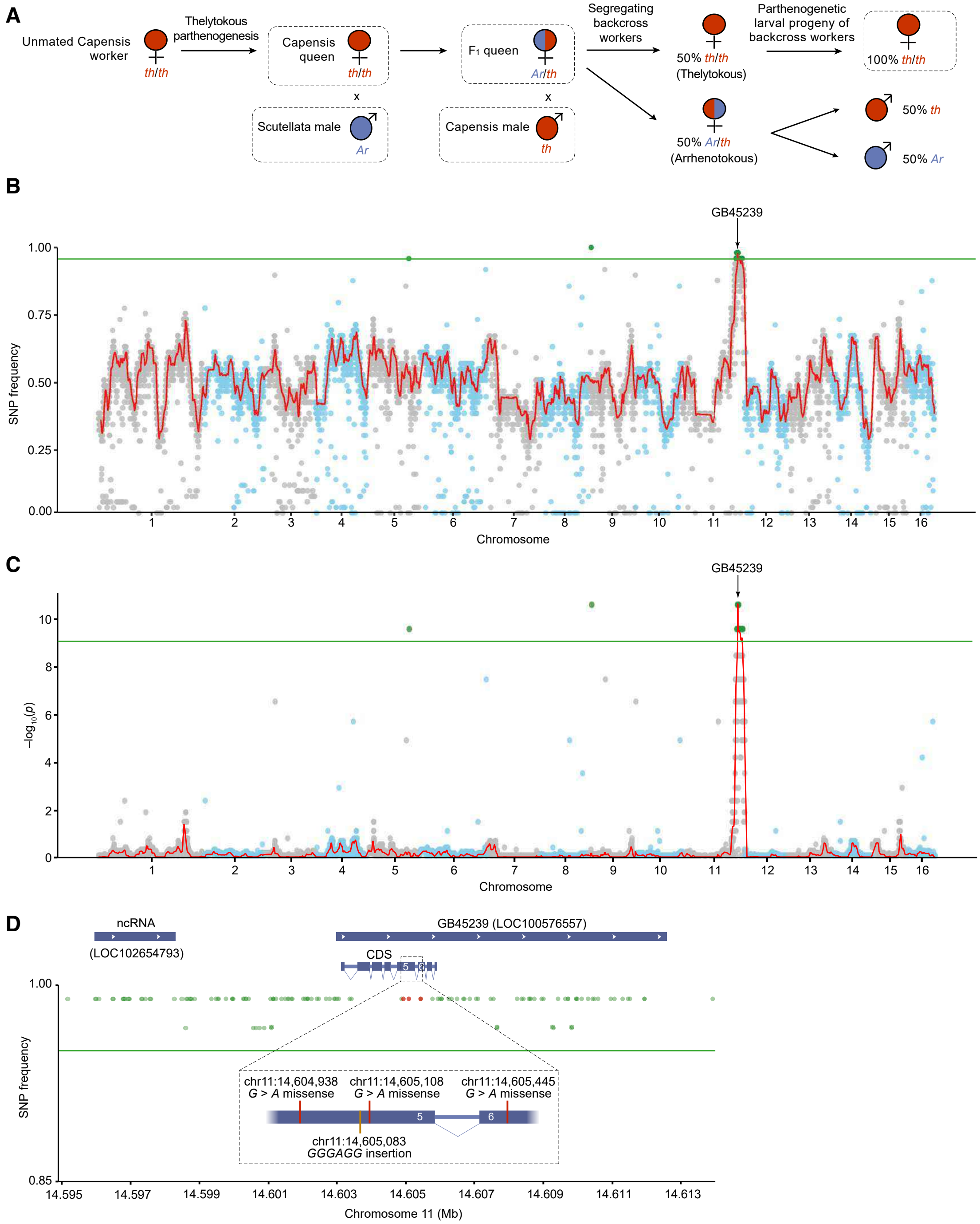

Figure 1. The Association between Genetic Variants and Thelytoky

(A) Crossing design. Individuals used for whole-genome sequencing are circled. Expected alleles are shown below each individual (red, recessive thelytokous *th*; blue, dominant arrhenotokous *Ar*).

(B) Frequency score of all informative SNPs along the genome. SNPs are represented by gray and blue dots on adjacent chromosomes. The red line is the average frequency score across 100 SNP sliding windows (25 SNP step size). The green line is the 95% cutoff used to identify HF SNPs (green dots).

(C) $-\log_{10}$(p value) of the frequency of all informative SNPs against the expected 0.5 (Benjamini-Hochberg [BH]-corrected binomial tests) along the genome. SNPs are represented by gray and blue dots on adjacent chromosomes. The red line is the average p value across 100 SNP sliding windows (25 SNP step size). The green line is the 99% percentile cutoff used to identify highly significant SNPs (green dots).

(D) Frequency score of all informative SNPs around gene GB45239 in chromosome 11. The green line represents the 95% cutoff used to identify HF SNPs (depicted in green). Non-synonymous variants are depicted in red. Enclosed is a close-up view of exons 5 and 6 encompassing the three missense variants and the insertion. CDS, coding sequence.

See also Figures S1, S2, and S5 and Tables S1, S2, and S4.

GB45239 Is a Putative Thelytoky-Causing Gene

The 23 kb region encompassing the top three windows intersected a non-coding RNA (ncRNA) (LOC102654793) and a protein-coding gene, GB45239 (uncharacterized LOC100576557) (Figures 1B–1D). There were 19 SNPs within the ncRNA, all of which were highly significant HF. There were 58 SNPs within GB45239, 55 of which were highly significant HF (average frequency score 0.976). Of particular interest were 8 highly significant HF SNPs located within the coding sequence: 3 intronic variants, 2 synonymous changes, and 3 non-synonymous changes (Figure 1D). All 3 non-synonymous changes were *G*-to-*A* substitutions, generating missense variants (Gly to Arg at position 361, Met to Ile at position 417, and Val to Ile at position 494) and were located in exons 5 and 6 (Figure 1D). GB45239 further had a *GGGAGG* (Gly and Arg at position 410) insertion located 26 bp upstream of the second missense variant in exon 5 (Figure 1D). Note that other transcribed regions, including protein-coding genes and ncRNAs, were present outside the top windows (Figure S2). However, none had non-synonymous substitutions, or similar densities of HF SNPs.

Male progeny of the backcross workers were expected to be 50% *Ar* (i.e., carrying the dominant arrhenotoky allele) and 50% *th* (Figures 1A and S1). We tested this prediction by sequencing a 624 bp product spanning the three missense variants in 41 male (haploid) offspring larvae of the backcross workers (which by definition were arrhenotokously produced). Twenty-one (51.22%) males had a *G*,*G*,*G* genotype (putatively *Ar*), whereas 20 (48.78%) males had an *A*,*A*,*A* genotype (putatively *th*). This ratio was not significantly different from 1:1 expectations (exact binomial test: $p > 0.999$). Note that any ratio greater than 2:1 would be significantly different ($p = 0.028$) from 1:1 with this sample size.

GB45239 Shows Capensis-Specific Polymorphisms

We screened an additional 443 genomes representing 18 different subspecies and/or lineages by using publicly available as well as new data to estimate the allelic frequencies of the polymorphisms identified in GB45239. If the missense variants are linked to *th*, then Capensis (thelytokous) bees should be homozygous *AA* at these loci, whereas non-Capensis (non-thelytokous) bees should be either homozygous *GG* or heterozygous *GA* at these loci (the *A* allele should be recessive). We found that almost all 109 Capensis bees examined were homozygous *AA* for the three missense variants (respectively, 100.00%, 99.08%, and 99.08%; Figure 2). Conversely, the *A* allele was absent from non-African bees (C, M, and O lineages; see [38]) and present at a low frequency in non-*capensis* African subspecies (A lineage) for the three loci (respectively, <4.77%, <4.45%, and <26.10%) (Figure 2). Furthermore, in the non-Capensis African subspecies *A. m. adansonii*, *A. m. monticola*, and *A. m. yemenitica*, when the *A* allele was present, all bees were heterozygous *GA* at these loci (and therefore non-thelytokous). Scutellata was the only exception, given that we observed a low frequency (4.35%) of bees being homozygous *AA* but only at the third variant (Figure 2). Capensis × Scutellata F_1 hybrids were all heterozygous *GA* at the three variants, as expected. The 6 bp insertion was found in all Capensis bees and was completely absent from any other bee, except for a few rare Scutellata and Africanized bees that were also heterozygous for the second missense variant (respectively, 4.44% and 1.64%) (Figure 2). The insertion appeared to be in perfect linkage disequilibrium with this missense variant. Overall, the three missense variants and the insertion formed one haplotype in over 99% of Capensis genomes. This haplotype was not observed in any other bee.

We also corroborated the finding that alleles associated with *th* are derived in Capensis by comparing patterns of genetic differentiation (pairwise F_{ST}) in seven African and European honeybee subspecies. We first detected SNPs with outlier F_{ST} values, defined here as SNPs with F_{ST} values above the 95% percentile in the genome-wide dataset for any given pairwise comparison. When we compared Capensis with the other six subspecies, our candidate gene GB45239 contained an unusually high number of outlier SNPs (18.17 ± 2.70 outlier SNPs, all $p < 0.00097$; Table S3). The three missense variants were outlier SNPs in each pairwise comparison. By contrast, the other six subspecies each had a lower number of outlier SNPs within GB45239 (between 6.83 ± 2.95 and 12.50 ± 3.19; between 1 and 3 significant pairwise comparisons; Table S3). Likewise, ncRNA LOC102654793, located ~5 kb upstream of GB45239, also showed a high number of outlier SNPs when we compared Capensis with the other six species (Capensis: 17.83 ± 1.14 outlier SNPs, all $p < 0.00001$; other species: between 4.00 ± 2.11 and 6.67 ± 2.96; between 1 and 3 significant pairwise comparisons). These analyses clearly indicate that derived mutations in Capensis are driving elevated patterns of genetic differentiation around GB45239 compared with that of other honeybee subspecies.

GB45239 Polymorphisms Segregate around South Africa's Hybrid Zone

We investigated the genotypes of 135 workers and 42 drones collected across South Africa at the three missense variants and the 6 bp insertion. Fifty-five workers and 19 drones were collected north of the hybrid zone and were predicted to be arrhenotokous Scutellata [22]. Eighty workers and 23 drones were collected south of the hybrid zone and were predicted to be thelytokous Capensis. The vast majority of bees south of the hybrid zone (94.81% of the workers and 100.00% of the drones; Figure 3) had a typical Capensis haplotype (i.e., homozygous *AA* for all three missense variants and the insertion present). By contrast, only 5.00% of the workers and 5.26% of the drones had such a haplotype north of the hybrid zone (Figure 3). Most of these northern bees had a wild-type haplotype (i.e., homozygous *GG* for all three missense variants and no insertion) as seen in most Scutellata bees (Figure 2), although there was more allelic variability at these sites (Figures 3B and 3C).

The fixation index indicated an overall deficit of heterozygotes caused by very high homozygosity for *A* in the south ($F = 0.453 \pm 0.134$; $F_{IS} = 0.307 \pm 0.038$). North and south populations had a very high degree of genetic differentiation ($F_{ST} = 0.640 \pm 0.037$). Polymorphisms in GB45239 therefore segregate very clearly between southern thelytokous bees and northern arrhenotokous bees.

GB45239 Encodes a Hymenopteran-Specific Protein Putatively Involved in Chromosome Segregation

Clear orthologs of GB45239 were identified in the genomes of sequenced *Apis* species (*A. cerana*, *A. dorsata*, and *A. florea*; 84.73%–88.36% amino acid identity), in other *Apidae* (e.g.,

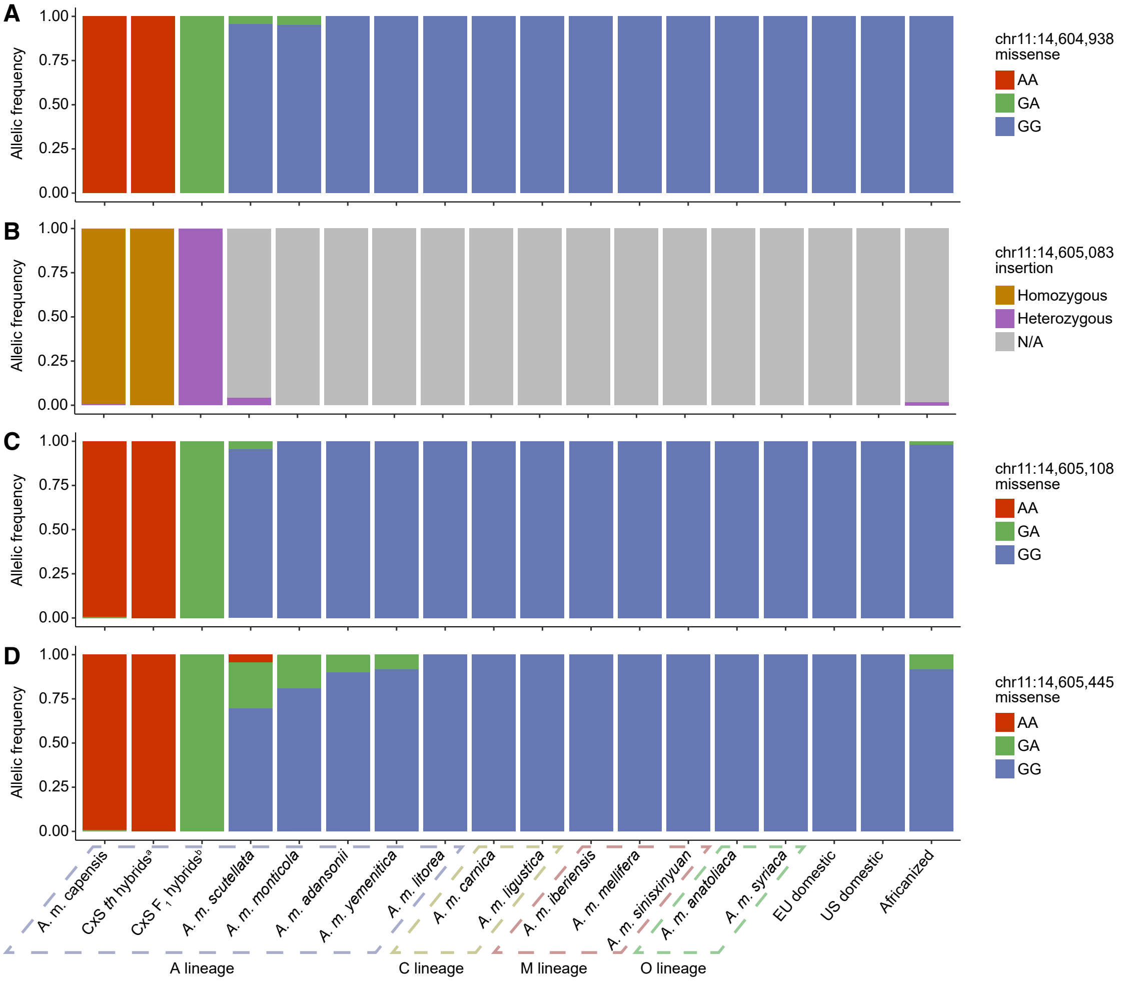

Figure 2. Allelic Frequency at Polymorphic Sites Identified in GB45239 across 18 Different Subspecies/Lineages

(A) Missense variant chr11:14,604,938.

(B) Six bp insertion chr11:14,605,083.

(C) Missense variant chr11:14,605,108.

(D) Missense variant chr11:14,605,445.

Subspecies are grouped by their evolutionary lineage [38]. [a]Female offspring of Capensis × Scutellata thelytokous backcross workers investigated in this study (excluding offspring #14); [b]F_1 female offspring of Capensis × Scutellata [39]; N/A, not applicable.

See also Tables S3 and S5.

bumblebees, 59.93%; stingless bees, 54.53%) and *Halictidae* (sweat bees, 44.88%), and in 14 ant species, one social and two parasitoid wasps (13.51%–29.05%) (Figures 4A and S3). No direct orthologs were evident in the genome of the sawfly *Cephus cinctus*, a basal lineage of Hymenoptera [40], nor in non-hymenopteran insects such as *Drosophila melanogaster* and *Bombyx mori*.

When considering the three missense variants associated with thelytoky in Capensis, the first and second sites were highly conserved in non-thelytokous honeybees, whereas the third site was more variable (Figure 4A). The 6 bp insertion was present in all other species examined apart from non-thelytokous *A. mellifera*, yet with different residues. These results suggest that GB45239 is evolutionarily derived in bees and further derived in Capensis. None of the ant species known to be capable of thelytokous reproduction and for which genomic data are available, i.e., *Ooceraea biroi*, *Wasmannia auropunctata* and *Vollenhovia emeryi* (reviewed in [26]), had the same residues at the three missense variants.

Using the NCBI conserved domain search, we found that GB45239's protein gave no significant hits to known domains that might hint at function. However, orthologs from other *Apidae* and ant species gave positive hits for a protein domain from the SMC (structural maintenance of chromosomes) superfamily. SMC proteins are present in all cellular organisms, and members of this family are involved in chromosome assembly, segregation, and sister chromatid adhesion, during mitosis and meiosis [41]. We used RaptorX [42] to predict *A. mellifera* GB45239's secondary and tertiary protein structures. GB45239's predicted structure consists of two domains (uGDT (GDT): 271 (45); p = 0.00003), N-terminal (residues 1–246) and C-terminal (residues 247–598; Figures 4B and 4C). Although the C-terminal domain is predicted to be mostly unstructured, the N-terminal domain has structural

Please cite this article in press as: Yagound et al., A Single Gene Causes Thelytokous Parthenogenesis, the Defining Feature of the Cape Honeybee *Apis mellifera capensis*, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.04.033

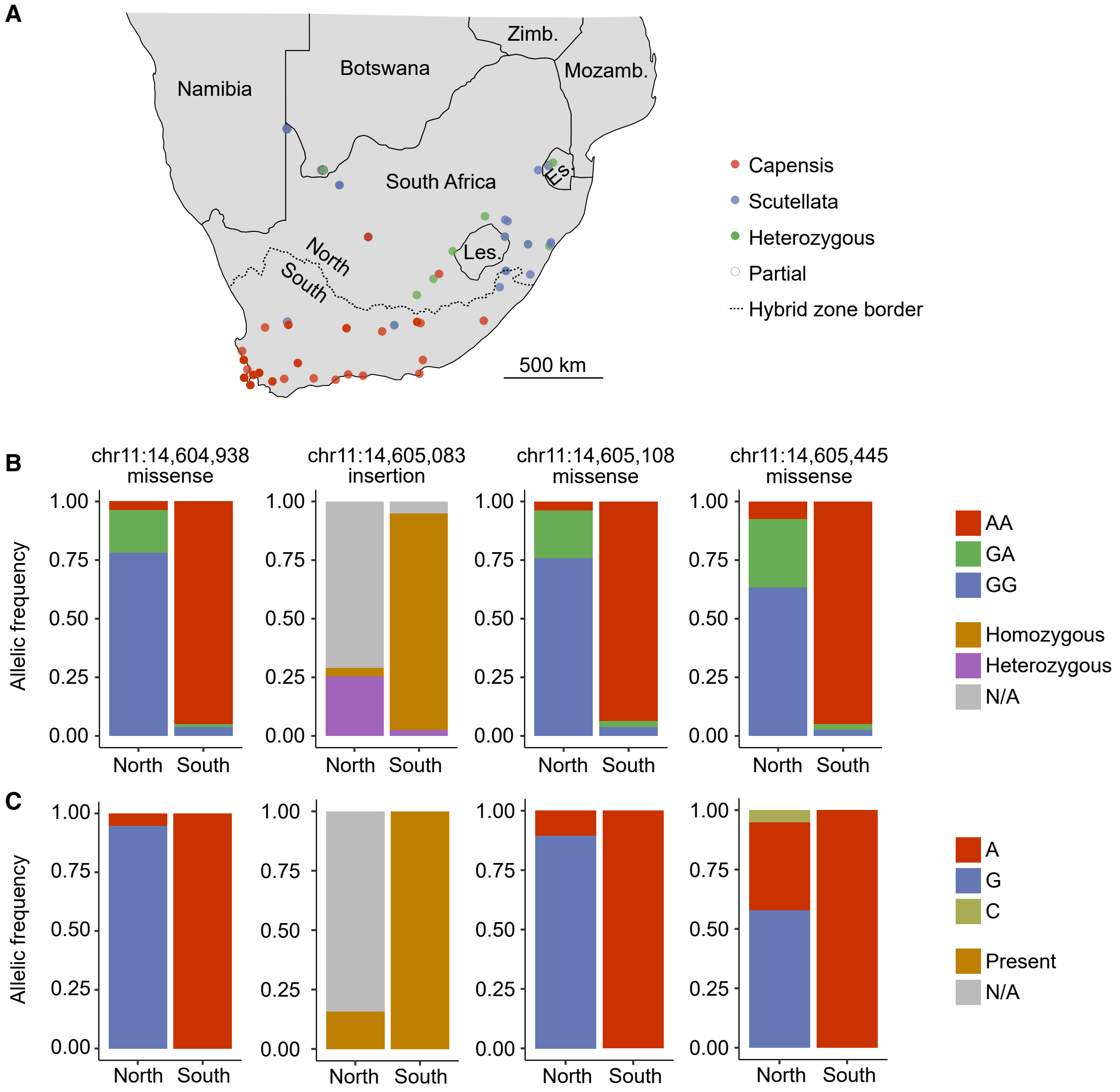

Figure 3. Segregation of GB45239 Genetic Variants around South Africa's Hybrid Zone

(A) Location and haplotype of worker samples from across Southern Africa. Capensis, homozygous *AA* at all missense variants and insertion present; Scutellata, homozygous *GG* at all missense variants and insertion absent; heterozygous, heterozygous at at least one site; partial, no data at at least one site.

(B and C) Allelic frequency at polymorphic sites identified in GB45239 for workers (B) and drones (C). N/A, not applicable.

hallmarks of an ABC ATPase head and a hinge dimerization domain connected by a coiled-coil arm (Figure 4B) that is structurally homologous to an SMC protein (PDB: 5XG2). GB45239 polymorphisms are all located in the C-terminal domain. Thus, although the overall predicted structure of GB45239 is similar in Capensis and arrhenotokous bees (Figure 4C), the functional consequences of the polymorphisms remain unknown. These results suggest that GB45239 is involved in meiosis. A failure of proper cytokinesis in meiosis II might differentiate thelytokous Capensis from arrhenotokous bees.

GB45239 Is Differentially Expressed in Capensis Ovaries

We expected GB45239 to be expressed in ovaries and/or early embryos if its role is linked to thelytoky. To test this prediction, we inspected all transcriptomic data mapping to GB45239 on NCBI. The strongest hit was for queen's ovary (SRA: PRJNA79571). We also retrieved transcriptomic data from [43] comparing gene expression levels in different tissues and castes across the honeybee's ontogeny. GB45239 was most highly expressed in the gaster of mated queens, most likely in the ovaries, and to some extent in early embryos (Figure S4). By comparison, expression levels were very low in the gaster of virgin queens and workers (all of which have inactive ovaries), and all other tissues.

Next, we investigated whether the GB45239 thelytokous allele was associated with changes in expression levels. We dissected backcross workers produced by a second F_1 queen that had been placed in queenless Scutellata micro colonies to induce egg laying. We expected half of these workers to be

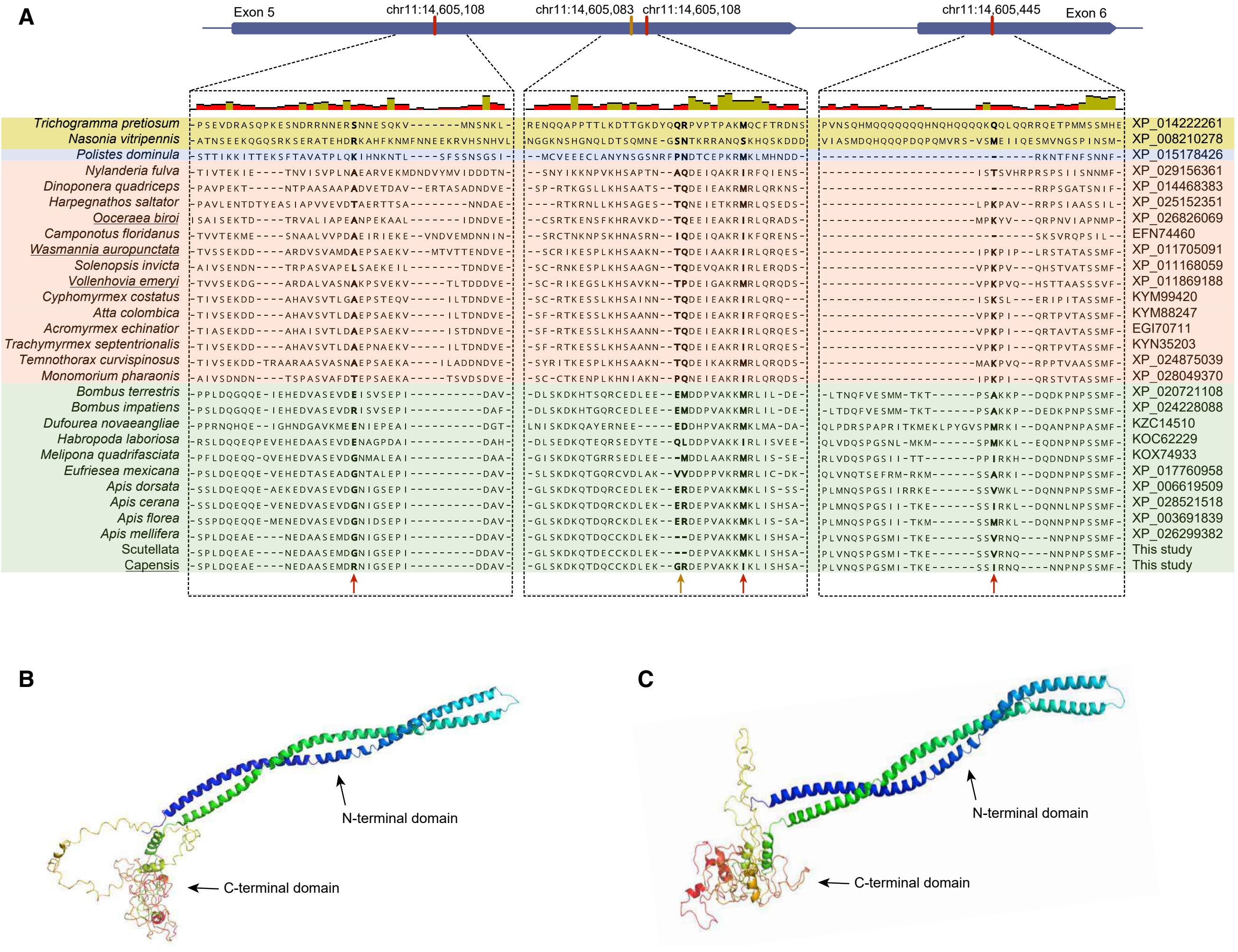
A
Exon 5
chr11:14,605,108
chr11:14,605,083
chr11:14,605,108
chr11:14,605,445
Exon 6
Trichogramma pretiosum XP_014222261
Nasonia vitripennis XP_008210278
Polistes dominula XP_015178426
Nylanderia fulva XP_029156361
Dinoponera quadriceps XP_014468383
Harpegnathos saltator XP_025152351
Ooceraea biroi XP_026826069
Camponotus floridanus EFN74460
Wasmannia auropunctata XP_011705091
Solenopsis invicta XP_011168059
Vollenhovia emeryi XP_011869188
Cyphomyrmex costatus KYM99420
Atta colombica KYM88247
Acromyrmex echinatior EGI70711
Trachymyrmex septentrionalis KYN35203
Temnothorax curvispinosus XP_024875039
Monomorium pharaonis XP_028049370
Bombus terrestris XP_020721108
Bombus impatiens XP_024228088
Dufourea novaeangliae KZC14510
Habropoda laboriosa KOC62229
Melipona quadrifasciata KOX74933
Eufriesea mexicana XP_017760958
Apis dorsata XP_006619509
Apis cerana XP_028521518
Apis florea XP_003691839
Apis mellifera XP_026299382
Scutellata This study
Capensis This study
B
N-terminal domain
C-terminal domain
C
N-terminal domain
C-terminal domain

Figure 4. Similarity of GB45239 Orthologs within Hymenopterans and Predicted Protein Structure

(A) Protein alignment of GB45239 orthologs across the Hymenoptera (parasitoid wasps, yellow; social wasps, blue; ants, red; bees, green) showing the position of the three missense variants (red arrows) and the 6 bp insertion (orange arrow). Species capable of thelytokous reproduction are underlined.

(B) Predicted protein structure of GB45239 for *A. mellifera*'s reference sequence.

(C) Predicted protein structure of GB45239 for Capensis.

See also Figure S3.

homozygous *th/th* (and therefore homozygous *AA* for the three missense variants) and the other half to be *Ar/th* (and therefore heterozygous *GA* for the three missense variants) (Figure 1A). RNA extracted from the workers' activated ovaries was used to perform qRT-PCR using primers spanning a 213 bp region across the polymorphic region in exons 5 and 6. We confirmed the genotype of the bees by sequencing a 624 bp product derived from genomic DNA spanning the polymorphic region. Twenty-three workers were homozygous *AA*, and 15 workers were heterozygous *GA*, not significantly different from Mendelian predictions (exact binomial test: $p = 0.256$). The expression level of GB45239 in *th/th* F_1 workers (presumably thelytokous) was significantly lower than in *Ar/th* F_1 workers (presumably arrhenotokous; Wilcoxon rank-sum test with continuity correction: $p < 0.00001$; Figure 5).

We searched for the presence of HF SNPs in the promoter region of GB45239 (defined as the 2 kb region upstream of the transcription start site (TSS) [44]). There were 25 polymorphisms in this region, including two HF SNPs in the 5′ UTR (Figure 1D). This suggests that polymorphisms in *cis*-regulatory modules might reduce expression of GB45239 in Capensis laying workers.

Previous Thelytoky Candidate Genes

GB45239 was the top candidate gene in a previous study that considered genomic differences between Capensis and other African subspecies [25]. None of the other candidate genes found in this study had any HF SNPs, and their average frequency score was around the genome average (Figure S5; Table S4). GB44980, also located on chromosome 11 only about 1 Mb upstream of GB45239, had a higher average frequency score (0.769) (Figure S5; Table S4). Yet it contained only intronic SNPs, and none of them were HF (maximum frequency score 0.776). Aumer et al.'s [36] candidate gene GB46427 located on

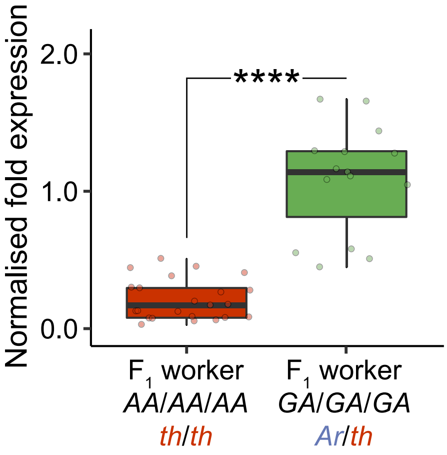

Figure 5. Expression Levels of GB45239 in the Ovary

qPCR normalized fold expression of GB45239 in the ovaries of (1) F_1 backcross workers homozygous for the thelytoky allele (i.e., homozygous for the 3 missense variants and the 6 bp insertion; red), and (2) F_1 backcross workers heterozygous for the thelytoky allele (i.e., heterozygous for the 3 missense variants and the 6 bp insertion; green). Boxplots represent median, interquartile range and 95% confidence interval. ****p < 0.00001 (Wilcoxon rank-sum test with continuity correction).

See also Figure S4.

chromosome 1 had no HF SNPs (average frequency score 0.569) (Figure S5; Table S4). Lattorff et al.'s [32] candidate gene GB48238 located on chromosome 13 also had no HF SNPs (average frequency score 0.388; Figure S5; Table S4). The lack of association between previous candidates and thelytoky was not caused by lack of informative SNPs (Table S4). Overall, these results indicate that, although GB45239 is associated with thelytoky in Capensis, all other candidates are not.

DISCUSSION

Our findings show that GB45239 (LOC100576557) on chromosome 11 causes thelytoky in Capensis and suggest that this gene has been central to the emergence of the Capensis phenotype. *th/th* bees are present with near 100% frequency south of the hybrid zone but are not present in any other honeybee population or species worldwide. GB45239 encodes a Hymenopteran-specific protein, derived in Capensis, and with a putative role in chromosome segregation. GB45239 is expressed in ovaries, and there was a downregulation in *th/th*-laying workers most likely associated with polymorphisms in the promoter region.

Our study rejects all previous *th* gene candidates, except for GB45239, as being the main determinants of thelytoky in Capensis. Wallberg et al. [25] found that Capensis bees differ from other bees in 12 regions on eight chromosomes. This analysis could not discriminate between loci that control thelytoky and loci that control other Capensis-specific traits (e.g., social parasitism, queen-like pheromones in workers, black body color), and so these loci most likely affect phenotypes that are associated with the Capensis phenotype but not with thelytoky. Strikingly, the gene that showed the most extreme genetic divergence in Capensis from other populations was the gene identified in our study (GB45239 on chromosome 11) [25]. This provides independent support to the association between GB45239 and thelytoky. However, we cannot rule out the possibility that additional genes affect thelytoky. Dominant variants or loci acting in epistasis could not be detected here and might play a role. Further, we are unable to say whether GB45239 affects other Capensis traits pleiotropically.

Thelytokous parthenogenesis differs from arrhenotokous parthenogenesis because of an unusual orientation of the meiotic spindle after meiosis II that results in the fusion of the two central pronuclei [12, 13]. GB45239 encodes a protein with a domain similar to SMC (structural maintenance of chromosomes) proteins. The putative function of GB45239, together with its expression in the ovaries gives additional support to its role in thelytokous parthenogenesis. Interestingly, SMC genes are evolutionarily labile [45]. GB45239 shows clear signs of genetic differentiation in Capensis, which could have facilitated the evolution of thelytoky in this lineage [14].

What is the molecular mechanism underlying the switch between arrhenotokous parthenogenesis in non-Capensis bees and thelytokous parthenogenesis in Capensis bees? Our results point toward two plausible hypotheses. First, we observed four polymorphisms within the protein-coding sequence of GB45239, which differentiate thelytokous and non-thelytokous bees. Missense mutations can impact protein structure and function [46] and can have dramatic phenotypic consequences [47]. These sequence polymorphisms could thus impact the function of GB45239's encoded protein in Capensis, potentially allowing the central fusion of two maternal pronuclei [13]. Second, we observed a downregulation of GB45239 in the ovaries of homozygous *th* laying workers (*AA* missense variants) compared with heterozygotes (*GA* missense variants). Changes in gene expression, potentially mediated by ncRNA LOC102654793, could thus be the difference between Capensis and non-Capensis bees, regardless of or in addition to the above-mentioned missense variants. We observed polymorphisms in the promoter region of GB45239, which could affect the binding of transcription factors in Capensis, resulting in lower levels of this transcript. This would imply that "normal" arrhenotokous parthenogenesis requires higher levels of gene expression, and that thelytokous parthenogenesis results from a "faulty" meiosis. This could explain why thelytoky occurs at very low frequency in non-Capensis bees [48, 49].

Patterns of genetic differentiation for thelytoky-linked alleles reflect a clear distinction between Capensis and non-Capensis honeybees. Homozygous (thelytokous) *th/th* individuals are virtually absent from non-Capensis bees, whereas heterozygous (phenotypically non-thelytokous) *Ar/th* individuals are found at low frequency in subspecies of the A (African) lineage and in Africanized bees. In South Africa, we found very limited introgression of the *Ar* allele south of the hybrid zone, and very limited introgression of *th/th* alleles north of the hybrid zone. These findings confirm the stability of the hybrid zone between Capensis and Scutellata populations [24, 50]. The reasons behind this stability are not fully elucidated. Several hypotheses have been suggested, including unfavorable dispersal benefits in the hybrid zone [18], adaptations to local biomes [14, 51], and selection against hybrids because of reduced fitness in genetically mixed colonies [22, 52]. The identification of robust genetic markers of *th* will now permit detailed examination of the dynamics of the

Please cite this article in press as: Yagound et al., A Single Gene Causes Thelytokous Parthenogenesis, the Defining Feature of the Cape Honeybee *Apis mellifera capensis*, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.04.033

hybrid zone and the fate of genotypically mixed colonies within the hybrid zone.

Identifying which gene controls thelytoky is an important step in understanding the evolution of this trait in Capensis, as well as in other thelytokous species. Although there seems to be no greater similarity between the protein sequence of GB45239 and its ortholog in thelytokous Hymenoptera compared with non-thelytokous Hymenoptera, there could be similarities in the regulation of gene expression among thelytokous species. Comparative transcriptomic analyses would be particularly valuable.

Conclusions

Many studies have sought to identify the genetic basis underlying changes in reproductive modes (e.g., [53, 54]), yet the exact gene(s) involved remain unknown. GB45239 is thus an important example of a characterized gene that switches an organism between sexual and asexual reproductive modes. Mutations in GB45239 and/or in *cis* cause thelytokous parthenogenesis by disrupting the final stages of meiosis. Thelytoky radically changes the kin structure of colonies [14, 26], releasing novel selective pressures that have resulted in the Capensis syndrome: highly reproductive workers, social parasitism, and social cancers.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Crosses
  - ○ Backcross Workers
  - ○ Sampling Progeny of Backcross Workers
  - ○ Whole Genome Sequencing
  - ○ Variant Calling
  - ○ Finding Regions Linked to Thelytoky
  - ○ Estimating the Number of Mother Workers
  - ○ SNPs Allelic Variability in Other Subspecies
  - ○ Patterns of Genetic Differentiation
  - ○ Transects
  - ○ GB45239's Hymenopteran Orthologs
  - ○ Gene Expression of Candidate Loci
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2020.04.033.

ACKNOWLEDGMENTS

This work was funded by the Australian Research Council project DP180101696 to B.P.O. and A.Z., the Natural Sciences and Engineering Research Council of Canada (NSERC), and a York University Research Chair in Genomics to A.Z. We thank Nick Smith for his assistance with variant calling. Three anonymous reviewers provided helpful comments on the manuscript.

AUTHOR CONTRIBUTIONS

B.P.O. and A.Z. designed research; B.P.O., B.Y., K.A.D., J.L., P.B., E.J.R., M.B., M.H.A., S.E.A., O.D., and G.B. performed research; B.Y., K.A.D., E.J.R., B.P.O., and A.Z. analyzed data; and B.Y., B.P.O., A.Z., and E.J.R. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 9, 2020
Revised: March 5, 2020
Accepted: April 15, 2020
Published: May 7, 2020

REFERENCES

1. Leonard, J.L. (2019). Transitions between Sexual Systems: Understanding the Mechanisms of, and Pathways between, Dioecy, Hermaphroditism and Other Sexual Systems (Springer).
2. Neiman, M., Sharbel, T.F., and Schwander, T. (2014). Genetic causes of transitions from sexual reproduction to asexuality in plants and animals. J. Evol. Biol. *27*, 1346–1359.
3. Otto, S.P. (2009). The evolutionary enigma of sex. Am. Nat. *174* (*Suppl 1*), S1–S14.
4. Williams, G.C. (1975). Sex and Evolution (Princeton University Press).
5. Maynard Smith, J. (1978). The Evolution of Sex (Cambridge University Press).
6. Bell, G. (1982). The Masterpiece of Nature: The Evolution and Genetics of Sexuality (University of California Press).
7. Winston, M.L. (1987). The Biology of the Honey Bee (Harvard University Press).
8. White, M.J.D. (1973). Animal Cytology and evOlution, Third Edition (Cambridge University Press).
9. Hepburn, H.R., and Crewe, R.M. (1991). Portrait of the Cape honeybee, *Apis mellifera capensis*. Apidologie (Celle) *22*, 567–580.
10. Ruttner, F. (1988). Biogeography and Taxonomy of Honeybees (Springer-Verlag).
11. Anderson, R.H. (1963). The laying worker in the Cape honeybee *Apis mellifera capensis*. J. Apic. Res. *2*, 85–92.
12. Verma, S., and Ruttner, F. (1983). Cytological analysis of the thelytokous parthenogenesis in the Cape honeybee (*Apis mellifera capensis* Escholtz). Apidologie (Celle) *14*, 41–57.
13. Cole-Clark, M.P., Barton, D.A., Allsopp, M.H., Beekman, M., Gloag, R.S., Wossler, T.C., Ronai, I., Smith, N., Reid, R.J., and Oldroyd, B.P. (2017). Cytogenetic basis of thelytoky in *Apis mellifera capensis*. Apidologie (Celle) *48*, 623–634.
14. Goudie, F., and Oldroyd, B.P. (2014). Thelytoky in the honey bee. Apidologie (Celle) *45*, 306–326.
15. Jordan, L.A., Allsopp, M.H., Oldroyd, B.P., Wossler, T.C., and Beekman, M. (2008). Cheating honeybee workers produce royal offspring. Proc. Biol. Sci. *275*, 345–351.
16. Holmes, M.J., Oldroyd, B.P., Allsopp, M.H., Lim, J., Wossler, T.C., and Beekman, M. (2010). Maternity of emergency queens in the Cape honey bee, *Apis mellifera capensis*. Mol. Ecol. *19*, 2792–2799.
17. Allsopp, M.H., Beekman, M., Gloag, R.S., and Oldroyd, B.P. (2010). Maternity of replacement queens in the thelytokous Cape honey bee *Apis mellifera capensis*. Behav. Ecol. Sociobiol. *64*, 567–574.

18. Neumann, P., Radloff, S.E., Moritz, R.F., Hepburn, H.R., and Reece, S.L. (2001). Social parasitism by honeybee workers (*Apis mellifera capensis* Escholtz): host finding and resistance of hybrid host colonies. Behav. Ecol. *12*, 419–428.
19. Moritz, R.F.A., Lattorff, H.M.G., Crous, K.L., and Hepburn, R.H. (2011). Social parasitism of queens and workers in the Cape honeybee (*Apis mellifera capensis*). Behav. Ecol. Sociobiol. *65*, 735–740.
20. Härtel, S., Neumann, P., Raassen, F.S., Moritz, R.F.A., and Hepburn, H.R. (2006). Social parasitism by Cape honeybee workers in colonies of their own subspecies (*Apis mellifera capensis* Esch). Ins. Soc. *53*, 183–193.
21. Beekman, M., Allsopp, M.H., Jordan, L.A., Lim, J., and Oldroyd, B.P. (2009). A quantitative study of worker reproduction in queenright colonies of the Cape honey bee, *Apis mellifera capensis*. Mol. Ecol. *18*, 2722–2727.
22. Beekman, M., Allsopp, M.H., Wossler, T.C., and Oldroyd, B.P. (2008). Factors affecting the dynamics of the honeybee (*Apis mellifera*) hybrid zone of South Africa. Heredity *100*, 13–18.
23. Hepburn, H.R., Jones, G.E., and Kirby, R. (1994). Introgression between *Apis mellifera capensis* Escholtz and *Apis mellifera scutellata* Lepeletier: The sting pheromones. Apidologie (Celle) *25*, 557–565.
24. Hepburn, H.R., Radloff, S.E., and Fuchs, S. (1998). Population structure and the interface between *Apis mellifera capensis* and *Apis mellifera scutellata*. Apidologie (Celle) *29*, 333–346.
25. Wallberg, A., Pirk, C.W., Allsopp, M.H., and Webster, M.T. (2016). Identification of multiple loci associated with social parasitism in honeybees. PLoS Genet. *12*, e1006097.
26. Goudie, F., and Oldroyd, B.P. (2018). The distribution of thelytoky, arrhenotoky and androgenesis among castes in the eusocial Hymenoptera. Ins. Soc. *65*, 5–16.
27. Neumann, P., and Hepburn, R. (2002). Behavioural basis for social parasitism of Cape honeybees (*Apis mellifera capensis*). Apidologie (Celle) *33*, 165–192.
28. Moritz, R.F.A., Simon, U.E., and Crewe, R.M. (2000). Pheromonal contest between honeybee workers (*Apis mellifera capensis*). Naturwissenschaften *87*, 395–397.
29. Neumann, P., Hepburn, H.R., and Radloff, S.E. (2000). Modes of worker reproduction, reproductive dominance and brood cell construction in queenless honeybee (*Apis mellifera* L.) colonies. Apidologie (Celle) *31*, 479–486.
30. Allsopp, M.H. (1992). The *Capensis* calamity. S. Afric. Bee J. *64*, 52–55.
31. Lattorff, H.M.G., Moritz, R.F.A., and Fuchs, S. (2005). A single locus determines thelytokous parthenogenesis of laying honeybee workers (*Apis mellifera capensis*). Heredity *94*, 533–537.
32. Lattorff, H.M.G., Moritz, R.F.A., Crewe, R.M., and Solignac, M. (2007). Control of reproductive dominance by the *thelytoky* gene in honeybees. Biol. Lett. *3*, 292–295.
33. Jarosch, A., Stolle, E., Crewe, R.M., and Moritz, R.F. (2011). Alternative splicing of a single transcription factor drives selfish reproductive behavior in honeybee workers (*Apis mellifera*). Proc. Natl. Acad. Sci. USA *108*, 15282–15287.
34. Chapman, N.C., Beekman, M., Allsopp, M.H., Rinderer, T.E., Lim, J., Oxley, P.R., and Oldroyd, B.P. (2015). Inheritance of thelytoky in the honey bee *Apis mellifera capensis*. Heredity *114*, 584–592.
35. Aumer, D., Allsopp, M.H., Lattorff, H.M.G., Moritz, R.F.A., and Jarosch-Perlow, A. (2017). Thelytoky in Cape honeybees (*Apis mellifera capensis*) is controlled by a single recessive locus. Apidologie (Celle) *48*, 401–410.
36. Aumer, D., Stolle, E., Allsopp, M., Mumoki, F., Pirk, C.W.W., and Moritz, R.F.A. (2019). A single SNP turns a social honey bee (*Apis mellifera*) worker into a selfish parasite. Mol. Biol. Evol. *36*, 516–526.
37. Christmas, M.J., Smith, N.M.A., Oldroyd, B.P., and Webster, M.T. (2019). Social parasitism in the honeybee (*Apis mellifera*) is not controlled by a single SNP. Mol. Biol. Evol. *36*, 1764–1767.
38. Franck, P., Garnery, L., Solignac, M., and Cornuet, J.M. (2000). Molecular confirmation of a fourth lineage in honeybees from the Near East. Apidologie (Celle) *31*, 167–180.
39. Smith, N.M.A., Yagound, B., Remnant, E.J., Foster, C.S.P., Buchmann, G., Allsopp, M.H., Kent, C.F., Zayed, A., Rose, S.A., Lo, K., et al. (2020). Paternally-biased gene expression follows kin-selected predictions in female honey bee embryos. Mol. Ecol. Published online March 27, 2020. https://doi.org/10.1111/mec.15419.
40. Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., et al. (2017). Evolutionary history of the Hymenoptera. Curr. Biol. *27*, 1013–1018.
41. Losada, A., and Hirano, T. (2005). Dynamic molecular linkers of the genome: the first decade of SMC proteins. Genes Dev. *19*, 1269–1287.
42. Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. Nat. Protoc. *7*, 1511–1522.
43. Warner, M.R., Qiu, L., Holmes, M.J., Mikheyev, A.S., and Linksvayer, T.A. (2019). Convergent eusocial evolution is based on a shared reproductive groundplan plus lineage-specific plastic genes. Nat. Commun. *10*, 2651.
44. Khamis, A.M., Hamilton, A.R., Medvedeva, Y.A., Alam, T., Alam, I., Essack, M., Umylny, B., Jankovic, B.R., Naeger, N.L., Suzuki, M., et al. (2015). Insights into the transcriptional architecture of behavioral plasticity in the honey bee *Apis mellifera*. Sci. Rep. *5*, 11136.
45. King, T.D., Leonard, C.J., Cooper, J.C., Nguyen, S., Joyce, E.F., and Phadnis, N. (2019). Recurrent losses and rapid evolution of the condensin II complex in insects. Mol. Biol. Evol. *36*, 2195–2204.
46. Studer, R.A., Dessailly, B.H., and Orengo, C.A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochem. J. *449*, 581–594.
47. Vitkup, D., Sander, C., and Church, G.M. (2003). The amino-acid mutational spectrum of human genetic disease. Genome Biol. *4*, R72.
48. Gloag, R., Remnant, E.J., and Oldroyd, B.P. (2019). The frequency of thelytokous parthenogenesis in *Apis mellifera ligustica* virgin queens. Apidologie (Celle) *50*, 295–303.
49. Tucker, K.W. (1958). Automictic parthenogenesis in the honey bee. Genetics *43*, 299–316.
50. Hepburn, H.R., and Radloff, S. (1998). Honeybees of Africa (Springer-Verlag).
51. Hepburn, H.R., and Jacot Guillarmod, A. (1991). The Cape honeybee and the fynbos biome. S. Afr. J. Sci. *87*, 70–73.
52. Beekman, M., Allsopp, M.H., Holmes, M.J., Lim, J., Noach-Pienaar, L.-A., Wossler, T.C., and Oldroyd, B.P. (2012). Racial mixing in South African honeybees: The effects of genotype mixing on reproductive traits of workers. Behav. Ecol. Sociobiol. *66*, 897–904.
53. Jaquiéry, J., Stoeckel, S., Larose, C., Nouhaud, P., Rispe, C., Mieuzet, L., Bonhomme, J., Mahéo, F., Legeai, F., Gauthier, J.P., et al. (2014). Genetic control of contagious asexuality in the pea aphid. PLoS Genet. *10*, e1004838.
54. Sandrock, C., and Vorburger, C. (2011). Single-locus recessive inheritance of asexual reproduction in a parasitoid wasp. Curr. Biol. *21*, 433–437.
55. Smith, N.M.A., Wade, C., Allsopp, M.H., Harpur, B.A., Zayed, A., Rose, S.A., Engelstädter, J., Chapman, N.C., Yagound, B., and Oldroyd, B.P. (2019). Strikingly high levels of heterozygosity despite 20 years of inbreeding in a clonal honey bee. J. Evol. Biol. *32*, 144–152.
56. Wallberg, A., Schöning, C., Webster, M.T., and Hasselmann, M. (2017). Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. PLoS Genet. *13*, e1006792.
57. Fuller, Z.L., Niño, E.L., Patch, H.M., Bedoya-Reina, O.C., Baumgarten, T., Muli, E., Mumoki, F., Ratan, A., McGraw, J., Frazier, M., et al. (2015). Genome-wide analysis of signatures of selection in populations of African honey bees (*Apis mellifera*) using new web-based tools. BMC Genomics *16*, 518.
58. Harpur, B.A., Kent, C.F., Molodtsova, D., Lebon, J.M., Alqarni, A.S., Owayss, A.A., and Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. Proc. Natl. Acad. Sci. USA *111*, 2614–2619.

59. Nelson, R.M., Wallberg, A., Simões, Z.L.P., Lawson, D.J., and Webster, M.T. (2017). Genomewide analysis of admixture and adaptation in the Africanized honeybee. Mol. Ecol. *26*, 3603–3617.

60. Liu, H., Zhang, X., Huang, J., Chen, J.Q., Tian, D., Hurst, L.D., and Yang, S. (2015). Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. Genome Biol. *16*, 15.

61. Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simões, Z.L., Allsopp, M.H., Kandemir, I., De la Rúa, P., et al. (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. Nat. Genet. *46*, 1081–1088.

62. Christmas, M.J., Wallberg, A., Bunikis, I., Olsson, A., Wallerman, O., and Webster, M.T. (2019). Chromosomal inversions associated with environmental adaptation in honeybees. Mol. Ecol. *28*, 1358–1374.

63. Chen, C., Liu, Z., Pan, Q., Chen, X., Wang, H., Guo, H., Liu, S., Lu, H., Tian, S., Li, R., and Shi, W. (2016). Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera sinisxinyuan* n. ssp. Mol. Biol. Evol. *33*, 1337–1348.

64. Kadri, S.M., Harpur, B.A., Orsi, R.O., and Zayed, A. (2016). A variant reference data set for the Africanized honeybee, *Apis mellifera*. Sci. Data *3*, 160097.

65. Ronai, I., Oldroyd, B.P., Barton, D.A., Cabanes, G., Lim, J., and Vergoz, V. (2016). Anarchy is a molecular signature of worker sterility in the honey bee. Mol. Biol. Evol. *33*, 134–142.

66. Scharlaken, B., de Graaf, D.C., Goossens, K., Brunain, M., Peelman, L.J., and Jacobs, F.J. (2008). Reference gene selection for insect expression studies using quantitative real-time PCR: The head of the honeybee, *Apis mellifera*, after a bacterial challenge. J. Insect Sci. *8*, 1–10.

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

68. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

69. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

70. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv, arXiv:1207.3907. https://arxiv.org/abs/1207.3907.

71. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

72. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.

73. R Core Team (2017). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

74. Sedlazeck, F.J., Rescheneder, P., and von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics *29*, 2790–2791.

75. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

76. Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. *103*, 338–348.

77. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

78. Peakall, R.O.D., and Smouse, P.E. (2006). Genalex 6: Genetic analysis in Excel. Population genetic software for teaching and research. Mol. Ecol. Notes *6*, 288–295.

79. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat. Protoc. *10*, 845–858.

80. Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

81. Ishida, T., and Kinoshita, K. (2007). PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res. *35*, W460–W464.

82. Harbo, J.R. (1986). Propagation and instrumental insemination. In Bee Genetics and Breeding, T.E. Rinderer, ed. (Academic Press), pp. 361–389.

83. Solignac, M., Vautrin, D., Loiseau, A., Mougel, F., Baudry, E., Estoup, A., Garnery, L., Haberl, M., and Cornuet, J.-M. (2003). Five hundred and fifty microsatellite markers for the study of the honeybee (*Apis mellifera* L.) genome. Mol. Ecol. Notes *3*, 307–311.

84. Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory Press).

85. Wallberg, A., Bunikis, I., Pettersson, O.V., Mosbech, M.-B., Childers, A.K., Evans, J.D., Mikheyev, A.S., Robertson, H.M., Robinson, G.E., and Webster, M.T. (2019). A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. BMC Genomics *20*, 275.

86. Harpur, B.A., Dey, A., Albert, J.R., Patel, S., Hines, H.M., Hasselmann, M., Packer, L., and Zayed, A. (2017). Queens and workers contribute differently to adaptive evolution in bumble bees and honey bees. Genome Biol. Evol. *9*, 2395–2402.

87. Jordan, L.A., Allsopp, M.H., Beekman, M., Wossler, T.C., and Oldroyd, B.P. (2008). Inheritance of traits associated with reproductive potential in Apis mellifera capensis and Apis mellifera scutellata workers. J. Hered. *99*, 376–381.

88. Wallberg, A., Glémin, S., and Webster, M.T. (2015). Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. PLoS Genet. *11*, e1005189.

89. Beye, M., Gattermeier, I., Hasselmann, M., Gempe, T., Schioett, M., Baines, J.F., Schlipalius, D., Mougel, F., Emore, C., Rueppell, O., et al. (2006). Exceptionally high levels of recombination across the honey bee genome. Genome Res. *16*, 1339–1344.

90. Solignac, M., Mougel, F., Vautrin, D., Monnerot, M., and Cornuet, J.-M. (2007). A third-generation microsatellite-based linkage map of the honey bee, *Apis mellifera*, and its comparison with the sequence-based physical map. Genome Biol. *8*, R66.

91. Harpur, B.A., Guarna, M.M., Huxter, E., Higo, H., Moon, K.-M., Hoover, S.E., Ibrahim, A., Melathopoulos, A.P., Desai, S., Currie, R.W., et al. (2019). Integrative genomics reveals the genetics and evolution of the honey bee's social immune system. Genome Biol. Evol. *11*, 937–948.

92. Walsh, P.S., Metzger, D.A., and Higuchi, R. (1991). Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques *10*, 506–513.

93. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.

94. Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. *39*, D225–D229.

95. Oldroyd, B.P., Halling, L.A., Good, G., Wattanachaiyingchareon, W., Barron, A.B., Nanork, P., Wongsiri, S., and Ratnieks, F.L.W. (2001). Worker policing and worker reproduction in *Apis cerana*. Behav. Ecol. Sociobiol. *50*, 371–377.

Please cite this article in press as: Yagound et al., A Single Gene Causes Thelytokous Parthenogenesis, the Defining Feature of the Cape Honeybee *Apis mellifera capensis*, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.04.033

96. Pfaffl, M.W., Tichopad, A., Prgomet, C., and Neuvians, T.P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper–Excel-based tool using pair-wise correlations. Biotechnol. Lett. *26*, 509–515.

97. Taylor, S.C., Nadeau, K., Abbasi, M., Lachance, C., Nguyen, M., and Fenrich, J. (2019). The ultimate qPCR experiment: Producing publication quality, reproducible data the first time. Trends Biotechnol. *37*, 761–774.

98. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Phenol: Chloroform: Isoamyl alcohol mixture	Sigma-Aldrich	Cat#7761; CAS: 136112-00-0
TRI Reagent®	Sigma-Aldrich	Cat#T9424; CAS: 108-952; CAS: 593-84-0
Taq-Ti	Fisher Biotech	Cat#TAQ-Ti
SuperScript III First-Strand Synthesis System	Thermo Fisher Scientific	Cat#18080051
Chelex® 100 Chelating Resin	Bio-rad	Cat#1432832
Critical Commercial Assays		
Qubit RNA BR Assay Kit	Thermo Fisher Scientific	Cat#Q10210
Sso Advanced Universal SYBR® Green Supermix	Bio-Rad	Cat#1725272
Deposited Data		
Raw data	This paper	SRA: PRJNA592197; PRJNA592273
A. m. capensis genomes	[37]	SRA: PRJNA521424
A. m. capensis genomes; *A. m. scutellata* genomes; Capensis x Scutellata hybrids genomes	[39]	SRA: PRJNA591427
A. m. capensis Clones genomes	[55]	SRA: PRJNA496560
A. m. capensis genomes	[36]	SRA: PRJNA507348
A. m. monticola genomes; *A. m. scutellata* genomes	[56]	SRA: PRJNA357367
A. m. yemenitica genomes; *A. m. litorea* genome; *A. m. monticola* genome; *A. m. scutellata* genomes	[57]	SRA: PRJNA237819
A. m. carnica genomes; *A. m. mellifera* genomes; *A. m. yemenitica* genomes; *A. m. iberiensis* genomes	[58]	SRA: PRJNA216922
Africanised bees genomes	[59]	SRA: PRJNA350769
A. m. ligustica genomes	[60]	SRA: PRJNA252997
A. m. adansonii genomes; *A. m. anatoliaca* genomes; *A. m. capensis* genomes; *A. m. carnica* genomes; *A. m. iberiensis* genomes; *A. m. ligustica* genomes; *A. m. mellifera* genomes; *A. m. scutellata* genomes; *A. m. syriaca* genomes; Africanised bees genomes; European domestic bees genomes; American domestic bees genomes	[61]	SRA: PRJNA236426
A. m. scutellata genomes	[62]	SRA: PRJNA481428
A. m. sinisxinyuan genomes	[63]	SRA: PRJNA301648
Africanised bees genomes	[64]	SRA: PRJNA324081
Gene expression data for GB45239 in different tissues, life stages and castes	[43]	Github: https://github.com/warnerm/devnetwork
Oligonucleotides		
GB45239 PCR forward primer: 5′-ACCACCATCCAATATTGAAGC-3′	This study	N/A

(*Continued on next page*)

Please cite this article in press as: Yagound et al., A Single Gene Causes Thelytokous Parthenogenesis, the Defining Feature of the Cape Honeybee *Apis mellifera capensis*, Current Biology (2020), https://doi.org/10.1016/j.cub.2020.04.033

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GB45239 PCR reverse primer: 5′-GCCATGTTCTCATCGATACAG-3′	This study	N/A
GB45239 qPCR forward primer: 5′-CCATTAGAGAGGACGAAGAGC-3′	This study	N/A
GB45239 qPCR reverse primer: 5′-GAACATACTGGAAGGATTAGG-3′	This study	N/A
Actin qPCR forward primer: 5′-TGCCAACACTGTCCTTTCTG-3′	[65]	N/A
Actin qPCR reverse primer: 5′-AGAATTGACCCACCAATCCA-3′	[65]	N/A
Ef1α qPCR forward primer: 5′-TGCAACCTACTAAGCCGATG-3′	[66]	N/A
Ef1α qPCR reverse primer: 5′-GACCTTGCCCTGGGTATCTT-3′	[66]	N/A
Software and Algorithms		
FastQC	Babraham Bioinformatics, UK	http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic	[67]
BWA	http://bio-bwa.sourceforge.net	[68]
SAMtools	http://samtools.sourceforge.net	[69]
Picard	N/A	http://broadinstitute.github.io/picard
FreeBayes	https://github.com/ekg/freebayes	[70]
vcflib	N/A	https://github.com/vcflib/vcflib#vcflib
VCFtools	https://vcftools.github.io/index.html	[71]
SnpEff	http://snpeff.sourceforge.net	[72]
R	https://www.r-project.org	[73]
NextGenMap	http://cibiv.github.io/NextGenMap/	[74]
GATK	https://gatk.broadinstitute.org/hc/en-us	[75]
Beagle	https://faculty.washington.edu/browning/beagle/beagle.html	[76]
Geneious	Biomatters, New Zealand	[77]
GenAlEx	https://biology-assets.anu.edu.au/GenAlEx/Welcome.html	[78]
Phyre2	http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index	[79]
RaptorX	http://raptorx.uchicago.edu	[42]
IUPred	https://iupred2a.elte.hu	[80]
PrDOS	http://prdos.hgc.jp/cgi-bin/top.cgi	[81]
Pipeline	https://github.com/Social-Insect-Genomics	N/A

LEAD CONTACT AND MATERIALS AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Boris Yagound (boris.yagound@sydney.edu.au).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The accession number for the whole-genome sequencing data reported in this paper is SRA: PRJNA592197. Pipeline used for analyses is available from Github (https://github.com/Social-Insect-Genomics).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We used two subspecies of the honeybee *Apis mellifera*. Arrhenotokous *A. m. scutellata* (Scutellata) bees were obtained from Douglas, Northern Cape (29°02′S, 23°45’E), well north of the hybrid zone. Thelytokous *A. m. capensis* (Capensis) bees were obtained from the Stellenbosch area, Western Cape (33°56’S, 18°52′E), well south of the hybrid zone.

For transects we collected single workers foraging on flowers in locations removed from known apiaries, or single bees from domestic but sedentary colonies throughout South Africa, Botswana and Eswatini over two years. We collected workers north of the hybrid zone in Aliwal North (30°42′S, 26°43’E), Ballito/Dolphin Coast (29°32′S, 31°13’E), Bergville KZN (28°43’S, 29°21E), Burgersdorp (30°59’S, 26°20’E), Champagne Castle (29°05’S, 29°20’E), Douglas (29°07’S, 23°45’E), Gege (26°58’S, 30°60’E), Gharagab camp (25°02′S, 20°04’E), Golden Gate Highlands Ntl Park (28°30’S, 28°36’E), Hilton (29°33’S, 30°17’E), Kokstad (30°30’S, 29°24’E), Mankayane (26°40’S, 31°03’E), Manzini (26°30’S, 31°22′E), Meerkat Manor (26°58’S, 21°49’E), Port Shepstone (30°43’S, 30°26’E), Shakaskraal (29°27’S, 31°13’E), Steynsburg (31°18’S, 25°49’E), Tswalu Nature Reserve (27°14’S, 22°24’E), Umdloti (29°40’S, 31°07’E), Vanstadensrus (29°59’S, 27°00’E), and Winterton (28°49’S, 29°31’E). We collected workers south of the hybrid zone in Aberdeen (32°48’S, 24°04’E), Addo Nlt Park Colchester Gate (32°24’S, 20°10’E), Buffeljagsrivier (34°12′S, 21°16’E), Caledon (34°23’S, 19°43’E), Cape Point (34°00’S, 18°63’E), Cederberg Wilderness Area (32°33’S, 19°12′E), Cradock (32°18’S, 25°65’E), Graaff-Reinet (32°25’S, 24°55’E), Grootvadersbosch (33°59’S, 20°49’E), Harkerville (34°03’S, 23°23’E), Idutywa (32°10’S, 28°31’E), Karoo Ntl Park (32°36’S, 22°54’E), Mossel Bay (32°24’S, 20°10’E), Mount Frere (30°89’S, 28°98’E), Mountain Zebra Ntl Park (32°14’S, 25°51’E), Mountain Zebra Ntl Park Craddock (33°48’S, 25°75’E), Port Elizabeth (33°96’S, 25°60’E), PPRI Stellenbosch (33°93’S, 18°87’E), Riviersonderend (34°36’S, 18°50’E), Robben Island (33°81’S, 18°37’E), Tankwa Ntl Park (32°14’S, 20°05’E), West Coast Ntl Park (33°17’S, 18°15’E), and Wilderness Ntl Park (33°99’S, 22°61’E). We collected drones north of the hybrid zone in Kwara (19°11’S, 23°27’E), Johannesburg (26°20’S, 28°05’E), Louis Trichardt (23°05’S, 29°90’E), and Molopo (25°72′S, 25°04’E). We collected drones south of the hybrid zone in Helderberg (34°05’S, 18°93’E), Kogelberg (34°29’S, 18°92′E), and Somerset West (34°08’S, 18°84’E).

METHOD DETAILS

Crosses

We used an arrhenotokous Scutellata queen as the mother of the drones used in our crosses. To obtain Capensis queens that were assuredly homozygous *th/th* we dequeened three Capensis colonies from the Stellenbosch area, and removed any developing queen cells one week later. The three colonies produced virgin queens from the progeny of the thelytokous workers that developed within the colonies. We instrumentally inseminated three of the Capensis queens, each with the semen of a different Scutellata male using standard methods [82]. We then reared six F_1 queens from the most populous colony, and inseminated each with the semen of a single Capensis male (see Figures 1A and S1 for the mating design). We retained the two grandparents and the fathering drones of our backcross colonies for sequencing.

Backcross Workers

We first selected the most vigorous backcross colony to provide individual workers to be phenotyped as being thelytokous or arrhenotokous. We attempted to use the method described in [35, 36] to classify individual workers as being thelytokous or arrhenotokous based on their progeny. We introduced individually marked day-old workers into micro-colonies containing c.a. 500 host workers of the Scutellata subspecies. (Scutellata workers do not generally reproduce in the presence of a Capensis worker that is actively laying [35],). The micro-colonies were housed in four screen tents, with approximately 15 colonies per tent. Colonies were regularly fed pollen and honey-icing sugar candy *ad libitum* as in Aumer et al. [35].

Unfortunately, we found that the majority of our backcross workers did not remain in their host colonies. If we clipped their wings so they could not fly, they got lost and died; if we did not clip their wings they often moved into other colonies. This behavior reduced our confidence that we could relate individual worker genotypes to their mode of parthenogenesis, and we abandoned the method. Nonetheless, we retained all backcross workers treated in this way and a sample of their brood under alcohol at −20°C.

As an alternative we selected two backcross colonies. The queens in these colonies had been laying for 10 weeks, more than sufficient time for complete replacement of the worker population. We removed the queens to induce oviposition by the backcross workers, and placed the colonies in isolation at least 500 m from any known colony and from each other, thereby reducing the frequency of reproductive parasitism by workers from other colonies. We retained the removed F_1 queens for sequencing and genotyping.

Sampling Progeny of Backcross Workers

To confirm sex and parentage of brood we genotyped the F_1 queen, the Scutellata male used for insemination, individual larvae and pupae of varying age using four microsatellite loci (A113, A14, A88 and B124 [83],). Brood carrying only hemizygous maternal alleles at all four loci was classified as being a haploid male laid by an arrhenotokous backcross worker (Figure 1A). Brood carrying a paternal allele at one or more loci (i.e., heterozygous) was classified as being a diploid female, a daughter of a thelytokous backcross worker. Brood that carried an allele not present in one or other parent at any locus were classified as the offspring of non-natal parasites. These individuals were not considered further. In one colony we identified both arrhenotokously produced male brood and thelytokously produced female brood, as well as some offspring of non-natal parasites. In the second colony most of the progeny were from non-natal parasites and we did not consider this colony further.

Whole Genome Sequencing

With our design, female offspring of backcross workers can be used to identify SNP loci that are linked to a putative locus that influences thelytoky (*th*) (Figure S1). SNP loci tightly linked to a thelytoky-causing locus will be homozygous for the SNP derived from Capensis parents of backcross workers in all female offspring (Figure S1). Loci that are unlinked to a thelytoky-causing locus will be heterozygous in half of backcross workers and their progeny, and homozygous for Capensis-derived SNPs in the other half (Figure S1). Males are less informative (Figure S1). DNA was isolated using standard phenol/chloroform/isoamyl alcohol extraction protocols [84]. Libraries were prepared (Nextera) from 49 female progeny of backcross workers which were sequenced on an Illumina NovaSeq 6000, S4 300 Cycle (one lane of 150 bp paired-end sequencing) at the Australian Genome Research facility, Melbourne.

Variant Calling

We checked the quality of the raw data with FastQC 0.11.7 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Trimmomatic 0.38 [67] was used to trim low quality reads and adaptor sequences with the following parameters: ILLUMINACLIP:/path/to/NexteraPE-PE.fa:2:30:10:8:TRUE HEADCROP:17 LEADING:22 TRAILING:22 SLIDINGWINDOW:4:22 MINLEN:36. Trimmed reads were mapped to the honeybee genome assembly Amel_HAv3.1 [85] using BWA 0.7.12 [68] with default parameters. Alignments were sorted and indexed with SAMtools 1.9 [69]. We marked PCR duplicates with Picard 1.119 (http://broadinstitute.github.io/picard). Single nucleotide polymorphisms (SNPs) were called using FreeBayes 1.2.0 [70]. SNPs were filtered with vcflib (https://github.com/vcflib/vcflib#vcflib). We excluded variants found within unplaced and non-nuclear regions and removed SNPs that had a quality score (QUAL) < 30, a mapping quality (MQM) < 40, a read depth (DP) < 10, a ratio of quality score to count of alternate allele (QUAL/AO) < 10, reads on one strand only (SAF = 0 or SAR = 0), less than one read on each side of the alternate allele (RPL < 1 or RPR < 1). We excluded SNPs with more than two alleles and those that fell within 10 bp of insertions or deletions using VCFtools 0.1.14 [71]. We further removed SNPs with a read depth falling outside 1.5 times the inter-quartile range [86]. We excluded variants falling within 5 bp of SNPs called heterozygous in drones when set as diploid [61]. We annotated SNPs using SnpEff 4.3 [72].

Finding Regions Linked to Thelytoky

Thelytoky is most likely influenced by recessive loci [10, 31–33, 87]. Therefore, a SNP that is in strong linkage disequilibrium with a locus that controls thelytoky should be homozygous in backcross workers (Figures 1A and S1). We thus removed all SNPs that were heterozygous in each backcross worker. We then concatenated the 49 workers' offspring VCF files into a single file using VCFtools. This produced a set of 1,529,462 homozygous SNPs for the workers' offspring.

SNPs linked with thelytoky-causing loci should be present and homozygous in the workers' Capensis grandmother and hemizygous in the Capensis father. These SNPs should also be heterozygous in the workers' F_1 mother. By contrast, SNPs linked to thelytoky-causing loci should not be present in the workers' Scutellata grandfather (Figure 1A). Following this reasoning, we sequentially excluded SNPs from the workers' offspring set to only retain SNPs that (1) were also present and hemizygous in the workers' Capensis father (1,084,071 SNPs left), (2) were present and homozygous in the workers' Capensis grandmother (462,888 SNPs left), (3) were present and heterozygous in the workers' F_1 mother (62,758 SNPs left), and (4) were absent in the workers' Scutellata grandfather. This left 62,526 SNPs that were potentially linked to thelytoky-causing loci.

SNPs linked to *th* should have a proportion close to 1 in thelytokously produced worker offspring. This proportion is predicted to be 0.5 at loci that are unlinked to *th* in the progeny of thelytokous laying workers (Figure S1). Focusing on the 62,526 SNPs that passed our filtering criteria, we calculated a frequency score for each SNP, defined as the proportion of workers' offspring being homozygous for the allele found in the Capensis father out of the total number (49) of workers' offspring. We then used a sliding window of 100 SNPs and a step size of 25 SNPs to identify putative regions associated with thelytoky. These regions would be characterized by a high density of SNPs with a very high frequency score (> 0.95, hereafter HF). We used SNPs-based windows instead of base pair-based windows to account for the uneven distribution of SNPs along the genome (meaning that some windows would have had a very high number of SNPs while others would have none). (In fact both approaches produced similar results).

Some authors have argued that more complex patterns of inheritance, including dominant loci, could underlie thelytoky at the genetic level [36]. We investigated this possibility by repeating the above-mentioned mapping procedures with all SNPs (either homozygous or heterozygous) found in the workers' offspring (total 1,862,632 SNPs). We used the same cutoffs to identify HF SNPs and highly significant SNPs. Not a single window was found above these cutoffs along the whole genome (maximum frequency score across all windows: 0.944, cutoff: 0.95; maximum $-\log_{10}$(p value) across all windows: 11.03, cutoff: 12.01; data not shown). We therefore report the data from the first set of analyses only in our Results.

Estimating the Number of Mother Workers

The precision of genomic mapping increases with the number of independent meiosis included in the analysis. To determine the number of independent mothers that contributed to the 49 backcross progeny used for mapping we examined their multi-locus SNP genotypes.

Consider a locus, A_1, that was heterozygous in the F_1 queen: $A_1^1A_1^2$. The male used to inseminate the queen may carry any of three alleles A_1^1, A_1^2 or A_1^3. Therefore worker progeny of the F_1 queen may be $A_1^1A_1^1$, $A_1^2A_1^2$, $A_1^1A_1^3$ or $A_1^2A_1^3$. Since the male transmits his allele to all his worker progeny, paternal alleles are uninformative, so we remove all paternal alleles. The simplified genotype of workers at informative loci is therefore A_1^1 or A_1^2 and their thelytokous progeny will be A_1^1 or A_1^2. (Some offspring will be $A_1^3A_1^3$ as a result of thelytokous recombination, but these locus/offspring combinations are not considered). Now consider a second unlinked locus, A_2, that is also heterozygous in the F_1 queen. The four possible multi-locus progeny are A_1^1-A_2^1, A_1^1-A_2^2 A_1^2-A_2^1 and A_1^2-A_2^2. The probability that the two-locus genotypes will be identical in daughters of the same worker is 1, and the probability that they will be the same in the progeny of different workers is $1/2^2$. The probability that progeny of the same worker would be declared as being the progeny of different workers in error is $1/2^n$, where n is the number of unlinked loci, heterozygous in the queen, that were considered.

To determine whether our mapping population were laid by different thelytokous workers we examined SNPs that were present in offspring (757,555-1,136,400 SNPs), as well as being heterozygous in the workers' F_1 mother (1,106,159 SNPs). Using R 3.3.3 [73], we then retrieved the SNPs that were unique to each worker's female offspring (i.e., not present in any of the other 48 offspring).

The recombination rate in honeybees, including African honeybees, is very high: 19-27 cM/Mb [60, 88–90] so that loci separated by more than 3 Mb are effectively unlinked. We used this threshold to consider any two loci in the same linkage group as being unlinked.

Our backcross progeny are supersisters ($r = 0.75$), and so the number of informative SNPs is low. The number of unlinked unique heterozygous SNPs per offspring was 12 on average (range 6-22). All 49 offspring had a statistically significant probability of having a unique mother worker (all $p < 0.0156$). For 28 of these offspring (57%), the probability of having a unique mother worker was highly significant (all $p < 0.00049$). Overall, it seems very likely that all or nearly all of the workers' female offspring analyzed originated from different workers.

SNPs that are tightly linked with *th* should have a frequency of 1 (Figure S1) in our mapping population of 49 workers. Most HF SNPs (91.39%) in the candidate region had a frequency of 0.980, i.e., they were present in all but one bee, offspring #14. Of 244 HF SNPs in this region, #14 had none, while all other 48 offspring had 242-244 HF SNPs. This bee lacked many SNPs present in the worker's Capensis father. This casts doubt on this worker's ancestry. For this reason, the average SNPs frequencies within the HF region was better estimated when excluding this individual. The HF region then had a frequency score of 0.998 across all 254 SNPs. The 8 HF SNPs located within the coding sequence of GB45239, including the 3 non-synonymous changes, had a frequency of 1.

SNPs Allelic Variability in Other Subspecies

We used publicly available sequences from NCBI's Sequence Read Archive to determine the allelic variability of the polymorphisms linked to *th*. We screened 443 genomes in addition to our genomes, representing 18 different subspecies/lineages. These were 16 *A. m. capensis* from South Africa (This study; SRA: PRJNA592273); 3 *A. m. capensis* from South Africa ([37]; SRA: PRJNA521424); 4 *A. m. capensis*, 5 *A. m. scutellata* and 4 Capensis x Scutellata hybrids from South Africa ([39]; SRA: PRJNA591427); 3 *A. m. capensis* Clones from South Africa ([55]; SRA: PRJNA496560); 71 *A. m. capensis* from South Africa ([36]; SRA: PRJNA507348); 20 *A. m. monticola* and 19 *A. m. scutellata* from Kenya ([56]; SRA: PRJNA357367); 2 *A. m. yemenitica*, 1 *A. m. litorea*, 1 *A. m. monticola* and 6 *A. m. scutellata* from Kenya ([57]; SRA: PRJNA237819); 9 *A. m. carnica* from Germany, Croatia and Slovenia, 5 *A. m. mellifera* from Poland, 10 *A. m. yemenitica* from Saudi Arabia and Yemen, 4 *A. m. iberiensis* from Spain ([58]; SRA: PRJNA216922); 22 Africanised bees from Brazil ([59]; SRA: PRJNA350769); 54 *A. m. ligustica* from China ([60]; SRA: PRJNA252997); 10 *A. m. adansonii* from Nigeria, 10 *A. m. anatoliaca* from Turkey, 10 *A. m. capensis* from South Africa, 10 *A. m. carnica* from Austria, 10 *A. m. iberiensis* from Spain, 10 *A. m. ligustica* from Italy, 20 *A. m. mellifera* from Norway and Sweden, 10 *A. m. scutellata* from South Africa, 10 *A. m. syriaca* from Jordan, 10 Africanised bees from Brazil, 20 European domestic bees from Sweden, 10 American domestic bees from the USA ([61]; SRA: PRJNA236426); 5 *A. m. scutellata* from Kenya ([62]; SRA: PRJNA481428); 10 *A. m. sinisxinyuan* from China ([63]; SRA: PRJNA301648); 29 Africanised bees from Brazil ([64]; SRA: PRJNA324081).

Patterns of Genetic Differentiation

Using a large number of publicly available genomes we assessed patterns of genetic differentiation in gene GB45239 and in ncRNA LOC102654793. We used 85 genomes representing seven different subspecies/lineages. Sequences were downloaded from NCBI's Sequence Read Archive, and included 20 *A. m. monticola*, 19 *A. m scutellata* ([56]; SRA: PRJNA357367), 10 *A. m sinisxinyuan* ([63]; SRA: PRJNA301648), 9 M lineage bees from Western Europe, 9 C lineage bees from Eastern Europe, 10 *A. m. yemenetica* ([58]; SRA: PRJNA216922), and 8 *A. m. capensis* ([36]; SRA: PRJNA507348 and [39]; SRA: PRJNA591427). Sequences were trimmed of adapters and low quality bases (< 20) using Trimmomatic 0.36 retaining reads greater than 50 bp in length. Trimmed reads were aligned to the honeybee genome assembly Amel_4.5 using NextGenMap 0.4.12 sequence aligner [74]. Resulting BAM files were sorted using SAMtools 1.3.1 and reads were marked for duplicates using Picard 2.1.0. Base quality scores were recalibrated using GATK 3.7 BaseRecalibrator and SNPs and indels were called using GATK 3.7 HaplotypeCaller [75]. We excluded variants located within five bp of an insertion or deletion, within five bp of areas of low complexity [91], and excluded variants from the unmapped scaffolds. Variants were also excluded using GATK VariantFiltration with the following thresholds: $MQ < 40.0$, $QD < 5.0$, $FS > 11.0$, MQRankSum $-2.0 < x > 2.0$, and ReadPosRankSum $-2.0 < x > 2.0$. Finally, Beagle 5.0 [76] was used to impute genotypes for bi-allelic loci whose subspecies/lineages had representative variant calls for at least 60% of individuals. Post filtering, 6,106,083 SNP loci remained for the analysis.

Transects

We collected single workers foraging on flowers in locations removed from known apiaries, or single bees from domestic but sedentary colonies throughout South Africa over two years. This collection allowed us to determine whether there is a relationship between putative thelytoky-causing alleles and location.

DNA was extracted from the hind legs of all samples using the Chelex® method [92]. PCRs were performed following standard conditions with custom primers covering the polymorphic region of GB45239 (F: ACCACCATCCAATATTGAAGC, R: GCCATGTTCTCATCGATACAG). PCR products were Sanger sequenced (Macrogen, Seoul) and the results were analyzed using Geneious [77]. We analyzed genetic data using GenAlEx 6 [78]. We determined genotype and allele frequencies at the three polymorphic sites for the populations north and south of the official hybrid zone 'border', i.e., the line that demarcates the magisterial districts through which bees must not be transported under South African Department of Agriculture regulations (Figure 3A). Weir and Cockerham's F_{ST} statistic [93] was used to measure the degree of population genetic differentiation between the area north (arrhenotokous Scutellata) and south (thelytokous Capensis) of the hybrid zone. F_{IS} was used to measure levels of homozygosity within the two populations. We calculated haplotype *F*-statistics by averaging values over loci for each population.

GB45239's Hymenopteran Orthologs

The protein coding region of GB45239 spans the first 8 exons resulting in a predicted 598-residue protein, followed by a 3′ region of 1.7kb non-coding RNA. We identified orthologs from sequenced insect genomes using reciprocal BLASTp searches to ensure one-to-one orthology of each top hit. Orthologs were aligned in Geneious [77] using Muscle, alignments were visually inspected and manually trimmed, and a phylogenetic tree was produced with the maximum likelihood method and 100 bootstrap replicates using the PhyML plugin in Geneious.

To examine GB45239 orthologs for homology to functionally characterized protein domains, we used the NCBI conserved domains search [94]. We first used Phyre2 [79] to generate a model of the tertiary protein structure of GB45239. Due to the absence of close homologs in the Protein Data Bank (PDB), Phyre2 predicted a model covering only ~6% of GB45239's sequence with low confidence. We then used RaptorX [42] which can predict secondary and tertiary protein structures without close homologs in PDB. IUPred [80] and PrDOS [81] were used to refine secondary structure predictions.

Gene Expression of Candidate Loci

We retrieved normalized gene expression data (Transcripts Per Million, TPM) for GB45239 from [43] in different tissues, life stages and castes on Github (https://github.com/warnerm/devnetwork).

We dissected the ovaries from 38 backcross workers produced by a second F_1 queen, a super sister to the mother of the mapping population. These workers had been individually laying in Scutellata micro colonies. Bees were stored frozen at −80°C and thawed briefly to dissect out ovaries. All had activated ovaries (white large ovaries with oocytes [95];). Total RNA was extracted from the ovary tissue from individual bees in TRI Reagent (Sigma Aldrich) using standard methods. RNA was quantified with a Qubit fluorometer (Life Technologies) and then diluted to 200 ng/μL. Genomic DNA was removed using TURBO DNase (Invitrogen). Ovary cDNA was synthesized using Superscript® III First-Strand Synthesis System for RT-PCR (Invitrogen). qRT-PCR was performed using custom primers (F: CCATTAGAGAGGACGAAGAGC; R: GAACATACTGGAAGGATTAGG) with SsoAdvanced Universal SYBR Green Supermix (Bio-Rad) in a CFX384 Real-Time System (Bio-Rad). Reference genes *Ef1α* and *Actin* were confirmed to be suitable to use for normalization using the software BestKeeper [96] with primers published in [66] and [65]. Gene expression level of the target gene GB45239 was normalized as in [97].

We determined the genotype of each bee by sequencing PCR products covering the polymorphic region of GB45239 using custom primers (F: ACCACCATCCAATATTGAAGC, R: GCCATGTTCTCATCGATACAG). Expressed transcripts were confirmed for the same region by performing RT-PCR on the ovary RNA and sequencing these products. PCR reactions were performed using TaqTi (Fisher Biotech). PCR products derived from genomic DNA as a template produced a 731 bp product. A 624 bp product confirmed the presence of GB45239 transcripts in the ovary tissue. Based on the sequencing information, the 38 backcross workers were split into two groups: 23 bees were determined to be homozygous *AA* for the three missense variants, and 15 bees were determined to be heterozygous *GA* for the three missense variants.

QUANTIFICATION AND STATISTICAL ANALYSIS

To estimate how SNPs deviated from frequencies expected for variants unlinked to *th*, we tested for each informative SNP if the observed frequency was significantly greater than 0.5 using binomial tests with R. *P*-values were corrected for each test using the Benjamini-Hochberg procedure [98]. We used the 99% percentile genome-wide as a cutoff to identify highly significant SNPs.

We used exact binomial tests with R to estimate i) if the ratio of *Ar* to *th* genotypes differed from the expected 1:1 in the male progeny of the backcross workers, and ii) if the ratio of *th/th* to *Ar/th* genotypes differed from the expected 1:1 in the backcross F_1 workers.

Pairwise patterns of genetic differentiation between subspecies/lineages were estimated with Weir and Cockerham's F_{ST} statistic [93] using VCFtools 0.1.17 (Table S5). The genome wide distribution of F_{ST} was estimated per SNP, and F_{ST} measures within the 0.95 quantile of the distribution were considered measures of genome outliers. Within the target genes GB45239 and LOC102654793, the number of variants falling within the 0.95 genome outlier range was calculated for each pairwise comparison. A Fisher exact test was used to compare the outlier distribution of GB45239 and LOC102654793 to the genome distribution using R. *P*-values, for each quantile comparison, were corrected for false discovery rate using the Benjamini-Hochberg correction [98].

We used a Wilcoxon rank sum test with continuity correction to compare the expression level of GB45239 in *th/th* and *Ar/th* F_1 workers.

Statistical details can be found in Results and in the figure legends.