The maintenance of polymorphism in an ancient social supergene

Jessica Purcell [1], German Lagunas-Robles [2], Christian Rabeling [3], Marek L. Borowiec [4, 5], Alan Brelsford [2*]

[1] Department of Entomology, University of California Riverside, 900 University Ave, Riverside, CA 92521, USA

[2] Department of Evolution, Ecology, and Organismal Biology, University of California Riverside, 900 University Ave, Riverside, CA 92521, USA

[3] School of Life Sciences, Arizona State University, 427 E Tyler Mall, Tempe, AZ 85281, USA

[4] Department of Entomology, Plant Pathology and Nematology, University of Idaho, 875 Perimeter Drive, Moscow, ID 83844, USA

[5] Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, 875 Perimeter Drive, Moscow, ID 83844, USA

Correspondence: Alan Brelsford, alanbe@ucr.edu

Running title: Polymorphism in an ancient supergene

Abstract

Supergenes, regions of the genome with suppressed recombination between sets of functional mutations, contribute to the evolution of complex phenotypes in diverse systems. Excluding sex chromosomes, most supergenes discovered so far appear to be young, being found in one species or a few closely related species. Here, we investigate how a chromosome harboring an ancient supergene has evolved over about 30 Ma. The *Formica* supergene underlies variation in colony queen number in at least five species. We expand previous analyses of sequence divergence on this chromosome to encompass about 90 species spanning the *Formica* phylogeny. Within the non-recombining region, the gene *knockout* contains 22 single nucleotide polymorphisms (SNPs) that are consistently differentiated between two alternative supergene haplotypes in divergent European *Formica* species, and we show that these same SNPs are present in most *Formica* clades. In these clades, including an early diverging Nearctic *Formica* clade, individuals with alternative genotypes at *knockout* also have higher differentiation in other portions of this chromosome. We identify hotspots of SNPs along this chromosome that are present in multiple *Formica* clades to detect genes that may have contributed to the emergence and maintenance of the genetic polymorphism. Finally, we infer the presence of three gene duplications on one haplotype, based on apparent heterozygosity within these genes in the genomes of haploid males. This study strengthens the evidence that this supergene originated early in the evolution of *Formica* and that just a few loci in this large region of suppressed recombination retain strongly differentiated alleles across contemporary *Formica* lineages.

Keywords: Formicinae, coadapted gene complex, suppressed recombination, gene flux

Introduction

Autosomal supergenes, sets of functional mutations linked in a region of suppressed or greatly reduced recombination, are associated with a variety of complex phenotypes in diverse eukaryotic lineages (Charlesworth 2016; Schwander et al. 2014; Thompson & Jiggins 2014). Supergenes underlie phenotypic polymorphisms including mating systems in fungi (Branco et al. 2018; Sun et al. 2019), reproductive strategies in plants (e.g. Huu et al. 2020, Li et al. 2016) and birds (e.g. Küpper et al. 2016; Lamichhaney et al. 2016; Tuttle et al. 2016), mimetic wing coloration in butterflies (Joron et al. 2011), and social organization in ants (Purcell et al. 2014; Wang et al. 2013). Known supergenes differ substantially in their age, with some present in a single species (e.g. ruffs, Küpper et al. 2016, Lamichhaney et al. 2016) and some spanning a few closely related species (e.g. in fire ants, Yan et al. 2020, Cohen & Privman 2020 and in *Heliconius* butterflies, Kronforst & Papa 2015). Just two autosomal supergenes so far may have an ancient origin: Ramanauskas & Igić (2017, 2021) suggested that the S-RNase-based self-incompatibility mechanism found in many flowering plants may have originated in the common ancestor of the core eudicots, and Brelsford et al. (2020) showed that the *Formica* ant social supergene, which underlies colony queen number and therefore controls polymorphic social organization, has persisted in species spanning an estimated 20-40 Ma of independent evolution.

Ancient supergenes provide an opportunity to investigate the evolutionary trajectory of a region of suppressed recombination across diverging species with different lifestyles. Brelsford et al. (2020) demonstrated that a large region of reduced recombination spans ~10.5 Mbp of a 14 Mbp chromosome in species from three distinct *Formica* clades. Within each of the well-studied species, numerous variants distinguish the two alternative supergene haplotypes (Brelsford et al. 2020). Surprisingly, only 142 single nucleotide polymorphisms (SNPs) were consistently

79 associated with alternative supergene haplotypes across these five species. The 142 trans-species
80 SNPs were distributed across the supergene region but no SNPs matching this pattern were found
81 elsewhere in the genome. These SNPs likely provide information about functionally important
82 regions of the supergene, because they would either be maintained in diverging lineages by
83 selection or rearrangements preventing recombination between alternative haplotypes. The
84 discovery of hotspots containing numerous trans-species, haplotype-specific SNPs that are
85 distributed across the supergene region is consistent with an analytical model investigating the
86 evolutionary fate of inversions with rare recombination or gene conversion events (Guerrero et
87 al. 2012). In this 'eroded strata' model (Brelsford et al. 2020), rare lineage-specific recombination
88 or gene conversion events would have homogenized alternative supergene haplotypes in
89 different lineages, but selection would have prevented the erosion of SNPs with functional
90 importance in each of the alternative haplotypes (Table 1 lays out the expectations of this model
91 at three timescales). In this study, we refer to variants that are found across divergent species as
92 trans-species, haplotype-specific SNPs.
93 *Formica* species frequently exhibit a polymorphism in colony queen number (e.g.
94 Chapuisat et al. 2004; DeHeer & Herbers 2004; Seppä et al. 2004; Bargum et al. 2007), with
95 some colonies having a single queen (= monogyne) and others containing multiple, often
96 unrelated, queens (= polygyne). These alternative social forms generally conform to the
97 "polygyny syndrome" where workers and queens are generally smaller, but colony size is larger,
98 in polygyne compared to monogyne colonies of the same species (Keller 1993; Rosset &
99 Chapuisat 2007). The discovery of a supergene associated with this variation has helped to shed
100 light on the proximate forces that maintain this phenotypic polymorphism (Purcell et al. 2014;
101 Avril et al. 2019, 2020). In *Formica selysi,* queens and workers from monogyne colonies

invariably have the Sm/Sm genotype, while individuals from polygyne colonies always have at least one copy of the Sp haplotype (diploids are either Sm/Sp or Sp/Sp; Purcell et al. 2014). Through high density linkage mapping, we have shown that recombination occurs across chromosome 3 in both homozygous forms, but recombination is suppressed in heterozygotes and there are a series of structural rearrangements between the two haplotypes (Purcell et al. 2014; Brelsford et al. 2020). More recently, a second linked supergene that is associated with split sex ratio phenotypes (i.e. whether colonies specialize in producing males or gynes) was identified in the North American species *Formica glacialis* and *Formica podzolica* (Lagunas-Robles et al. 2021).

In this study, we provide a comprehensive look at the evolutionary history of the chromosome carrying the *Formica* supergene in a comparison of over 90 species spanning the *Formica* tree of life. The dataset presented by Brelsford et al. (2020) did not include an early-diverging clade sister to most other *Formica* species (Romiguier et al. 2018, Borowiec et al. 2021), and included only European species. While *Formica* has received less research attention in the western hemisphere, North America is a hotspot of diversity in this genus. Some speciose North American clades are species-poor or absent in Europe (Borowiec et al. 2021). Our sample includes 11 genomes from the *neogagates-pallidefulva* clade, an early-diverging clade that is exclusively distributed in the New World, as well as several outgroup species, allowing us to examine the extent of trans-species variation on this chromosome across the entire genus. For 1-3 individuals per species, we assess whole genome sequence data including 90 new whole genome sequences that were generated for this study. Whole genome data is necessary, because we know that the clusters of trans-species, haplotype-specific SNPs are relatively small and likely to be missed with reduced representation sequencing. In this assessment, we maximize our

coverage of species and clades of *Formica*, but we do not yet know whether all of the species included in our analysis exhibit a polymorphism in colony queen number or sex ratio. We address the following questions: 1) Brelsford et al. (2020) found that the gene *knockout* harbored the only cluster of SNPs that were consistently found in 15 European species. We ask whether *knockout* remains a hotspot of trans-species, haplotype-specific SNPs across the whole genus. 2) Are there sets of divergent SNPs along chromosome 3 in members of the *neogagates-pallidefulva* clade, which is sister to all other *Formica* species in our analysis? If so, this would suggest that these genomic variants evolved prior to the diversification of the genus *Formica*. 3) Using our comparative approach, can we identify additional candidate genes that are consistently associated with alternative *knockout* variants in multiple *Formica* clades? Such genes may have contributed to selection for suppressed recombination to preserve favorable combinations of alleles in two or more genes and may continue to contribute to contemporary phenotypic polymorphisms. Genomic regions containing candidate genes would be characterized by trans-species, haplotype-specific SNPs in most *Formica* clades. These analyses expand our understanding of the evolutionary trajectory of the chromosome that harbors the *Formica* supergene and identify additional candidate genes therein.

Methods

Overview of samples included in the analysis

We obtained sequences from 114 *Formica* samples (1-3 individuals per species, most samples identified to species by Borowiec et al. 2021) and three outgroup samples (one *Iberoformica* and two *Polyergus;* Tables 2 and S1). Many samples (n = 80) were collected for a

concurrent study investigating the phylogenetic history of the genus *Formica* (Borowiec et al. 2021), 21 were previously assessed by Brelsford et al. (2020), and the remaining 16 were collected by Purcell and Brelsford (Table S1). We removed six samples from further consideration due to low sequence coverage and high levels of missing data (Table S1). Represented species spanned the phylogeny of the genus, including 10 species (11 individuals) from the *neogagates-pallidefulva* clade, as well as species from seven additional clades endemic to North America, including the *difficilis* and *integra* clades (Figure 1). Previous karyotyping of *Formica* species indicates that haploid chromosome number is generally 27 and 26 for non-parasitic and parasitic species, respectively (Hung 1969, Hauschteck-Jungen & Jungen 1976, Rosengren et al. 1980). The samples analyzed herein focused on maximizing the number of taxa represented, so the social structure of the source colony was not assessed, and the genotype of each individual at the supergene was not predicted in advance. As a result, we focus on assessing the sequence variation of these samples, and we interpret the supergene status with caution.

Whole genome sequencing

For 96 new samples, we extracted genomic DNA using Qiagen DNEasy Blood and Tissue kits. For 81 samples, we constructed whole genome libraries using KAPA library preparation kits (n=80) or a Nextera-based library protocol (Henderson and Brelsford 2021; n=1). These samples were first assigned to species based on morphological assessment (Table S1). Fifteen additional samples were sent to the UC Berkeley Vincent Coates Genome Sequencing Laboratory for library preparation (Tables 2, S1). All sequencing was performed by the UC Berkeley sequencing core, using the HiSeq 4000 platform with 150 bp paired-end reads.

Bioinformatics

We merged overlapping pair-end reads using PEAR (v0.9.10; Zhang et al. 2014). We aligned the reads to the published *F. selysi* chromosome-level genome assembly, which is based on individuals of monogyne origin (Sm haplotype; Brelsford et al. 2020) using BWA-MEM (v0.7.17; Li 2013) and removed PCR duplicates using Samtools (v1.8, Li et al. 2009). We called variants with Samtools mpileup, using only reads with a mapping quality of at least 20. We filtered the resulting variants using VCFtools (Danecek et al. 2011), retaining genotypes with minimum depth 2 and loci with missing data in < 50% of individuals. To identify groups of closely related species within our dataset, we generated a distance matrix based on genome-wide variants excluding chromosome 3 using Plink (v1.90b3.38, Purcell et al. 2007), and constructed a neighbor-joining tree using T-REX (Boc et al. 2012).

Assessing genotype at knockout

The gene *knockout* contained 22 trans-species, haplotype-specific SNPs that were consistently associated with queen number variation in five European *Formica* species (Brelsford et al. 2020). We examined the SNPs in this candidate gene to initially assess genotype variation in the newly sequenced species. We used VCFtools to export the genotypes of these 22 SNPs in all individuals and observed that most were either homozygous for the reference allele or were heterozygous.

Identifying additional trans-species, haplotype-specific SNPs

We then identified trans-species, haplotype-specific SNPs between the *knockout* homozygotes and heterozygotes of six clades of *Formica* along chromosome 3. In addition to

194 focusing on the early diverging *neogagates-pallidefulva* clade, we examined five other clades
195 that included at least four individuals with each of the alternative multi-locus genotypes at
196 *knockout*, including *fusca* #3, *fusca* #4, and *fusca* #5 clades, as well as two larger clades, one
197 containing all socially parasitic species and the other containing *fusca* clades #3, #4, and #5. We
198 used the --hardy function of VCFtools to count the number of individuals homozygous for the
199 reference allele, homozygous for the alternate allele, and heterozygous at each locus, separately
200 for *knockout* homozygotes and heterozygotes. We then used a custom script to identify SNPs that
201 were heterozygous in all genotyped *knockout* heterozygotes, and homozygous for the same allele
202 (either reference or alternate) in all genotyped *knockout* homozygotes, retaining only SNPs that
203 were genotyped in at least three heterozygous and three homozygous individuals in the focal
204 clade. In the *neogagates-pallidefulva* clade, we searched for these SNPs genome-wide. Since the
205 majority were found on chromosome 3, we plotted the number of SNPs meeting these criteria in
206 1 kbp windows along chromosome 3 for all comparisons. We identified positions along the
207 chromosome that consistently contained multiple trans-species, haplotype-specific SNPs in most
208 or all of the six clades. These hotspots of trans-species differentiation are more likely to harbor
209 genetic variants important in the origin and persistence of the supergene polymorphism. In
210 addition, we compared the trans-species, haplotype-specific SNPs identified in the *neogagates-*
211 *pallidefulva* clade across *fusca* clade #3, *exsecta* clade, and *rufa* clade, which were previously
212 examined by Brelsford et al. (2020) and contain samples of known social origin and supergene
213 status.
214
215 *Genes overlapping hotspots of differentiation between alternative haplotypes*

For four regions with high concentrations of trans-species, haplotype-specific SNPs, we used the annotated genome of *F. exsecta* (Dhaygude et al. 2019) to identify overlapping genes. We extracted the *F. selysi* reference genome sequence for each of these regions and aligned these sequences to the *F. exsecta* genome using BLAST (Altschul et al. 1997). We then scanned the *F. exsecta* annotation for genes overlapping these regions, downloaded the coding sequences of these genes, and aligned them back to the *F. selysi* reference to obtain the approximate locations of exons in *F. selysi*.

Results

Are knockout *SNPs polymorphic within other* Formica *clades?*

Variants spanning 8-22 SNPs in the gene *knockout* were detected in 11 out of 14 *Formica* clades examined in this dataset, including the *neogagates-pallidefulva* clade (Fig. 1). This comparison included 8 clades that were not represented in a previous trans-species analysis of the *Formica* supergene (Brelsford et al. 2020). In most of these new clades (6 out of 8), we detected individuals that were either homozygous for the *knockout* reference allele (based on alignment to the *F. selysi* reference genome, which is assembled from an individual of monogyne origin with the 'Sm' haplotype) or heterozygous. In the remaining two new clades, the *difficilis* and *dakotensis* clades, we detected only individuals homozygous at the reference allele (Fig. 1).

Since the SNPs distinguishing these groups are shared across divergent species, we infer that these genotypes are homologous to those previously identified in individuals of monogyne or polygyne origin (Brelsford et al. 2020, Fig. 1), although we cannot link alternative genotypes to phenotypes for most species in this dataset. For consistency within this study and across related studies, we refer to the three *knockout* multi-locus genotypes as 'Sm/Sm' (homozygous,

239 alleles match reference genome), 'Sm/Sp' (heterozygous), or 'Sp/Sp' (homozygous for the
240 alternative allele). Notably, we detected very few individuals that appeared Sp/Sp at *knockout*
241 (again, we cannot infer an association with polygyny based on current data). Overall, we
242 identified five Sp/Sp workers, plus three Sp males. All but one of these individuals were
243 selected for sequencing to represent the Sp haplotype in a previous study (Brelsford et al. 2020).
244 In the present study, we successfully sequenced 86 new *Formica* workers and two new males
245 without prior knowledge of their colony social structure. Of these, 63 workers were Sm/Sm at
246 *knockout*, 22 were Sm/Sp, and one was putatively Sp/Sp (Table S1, Fig. 1). Both males had the
247 Sm haplotype at *knockout*.
248 *Were haplotype-specific SNPs present in the common ancestor of extant* Formica*?*
249 Based on their genotypes at the gene *knockout*, our sample of *neogagates-pallidefulva*
250 clade members included five Sm/Sp heterozygotes and six Sm/Sm homozygotes. Since this
251 Nearctic clade diverged from other *Formica* clades very early, approximately 26 Ma ago, we
252 inferred that haplotype-specific SNPs were present in the common ancestor of the genus
253 *Formica.* Although we only sequenced a single *Iberoformica subrufa* individual, 19 of the 20
254 aligned *knockout* SNPs also matched the reference allele, suggesting that the reference haplotype
255 was already present in the most recent common ancestor of *Formica* and *Iberoformica*. Both
256 *Polyergus* samples exhibited a mix of reference and alternative alleles across this region.
257 Interestingly, some individuals in the *neogagates-pallidefulva* clade may carry a third
258 haplotype at the *knockout* locus. These individuals exhibited the Sm/Sm genotype at positions
259 11,910,116-11,910,323 bp and the Sm/Sp genotype at positions 11,910,356-11,910,880 bp
260 within our panel of 22 SNPs (Fig. 1). We considered individuals with a majority of Sm/Sp
261 genotypes from 11,910,356-11,910,880 bp, including both *F. incerta* individuals*, F. pallidefulva,*

F. lasioides, and *F. obtusopilosa,* to be heterozygous for subsequent analyses. By comparing the *knockout* heterozygotes and homozygotes from this clade, we identified trans-species, haplotype-specific SNPs specific to the *neogagates-pallidefulva* clade. This analysis yielded a total of 429 SNPs on chromosome 3. We found an additional 393 trans-species SNPs distributed in the rest of the genome, with the number of such SNPs per chromosome ranging from 1 on chromosome 18 to 69 on chromosome 17. With the exception of 2 SNPs that were found in the region from 0-2 Mbp, the SNPs on chromosome 3 spanned the 10.5 Mbp region that harbors the social supergene in other *Formica* species (Fig. 2, Purcell et al. 2014; Brelsford et al. 2020). In addition to the peak of differentiation that overlaps with *knockout* (at 11.9 Mbp), there are notable peaks harboring large numbers of SNPs at 2.3, 11.6, and 12.4 Mbp (Fig. 2A).

We assessed the number and position of trans-species, haplotype-specific SNPs along chromosome 3 within members of the *neogagates-pallidefulva* clade (Fig. 2A), within the socially parasitic *Formica* species, including the *exsecta, sanguinea*, *dakotensis*, *rufa, integra*, and *difficilis* clades (Fig. 2B), within the well-sampled *fusca* clades #3, 4, and 5 (Fig. 2C), and within each of these three *fusca* clades, #3 (Fig. 2D), #4 (Fig. 2E), and #5 (Fig. 2F). Across these comparisons, several small areas harboring numerous trans-species, haplotype-specific SNPs were consistently found at 2.3, 11.6, and 11.9 Mbp (the latter includes *knockout*, which was used to diagnose haplotype; Figs. 1, 2). We also observe peaks of differentiation at 12.4 Mbp in the *neogagates-pallidefulva* clade and in *fusca* clade #3.

Were additional candidate genes associated with alternative knockout *variants in multiple clades?*

We assessed the genes in the neighborhood of trans-species, haplotype-specific SNPs detected in the *neogagates-pallidefulva* clade. In addition to the gene *knockout*, we identified six other genes localized at or near one of four peaks of divergences between putative Sm/Sm and Sm/Sp individuals in this clade (Fig. 2A, Fig. 3). Like *knockout*, several of these additional genes are involved in neural development in *Drosophila* (Table 3) and may have contributed to the origin of the supergene.

To determine whether alternative alleles at these genes are likely to be maintained by selection in other *Formica* species, we then investigated the *neogagates-pallidefulva* clade trans-species, haplotype-specific SNPs in the workers and males from *exsecta*, *rufa,* and *fusca* clade #3 that include individuals with known monogyne or polygyne backgrounds, as described by Brelsford et al. (2020; see Table S1 and Fig. 1). For two of the four peaks of differentiation, at 2.3 Mbp and 12.4 Mbp along chromosome 3, we found that the haploid Sp males (and Sp/Sp homozygotes) appeared to be heterozygous (Fig. 3). This finding suggests a possible gene duplication event in the Sp haplotype: haploid individuals cannot be heterozygous, so the apparent heterozygosity we observe in Sp males must result from these individuals carrying two divergent copies of these regions of the genome. At an additional peak of differentiation at 11.6 Mbp within the supergene, we observed that only the *fusca* clade #3 Sp male and Sp/Sp workers appeared to be heterozygous, suggesting a possible lineage-specific duplication. Intriguingly, all of these clusters of differentiation overlap with annotated genes (Table 3). In the case of *single-minded*, all of the variants that differ between the two putative copies of the gene are found in introns (Fig. 3B). In contrast, the variants we detected in *AmGR10* include 100 SNPs in the introns and 32 in exons (Fig. 3C).

Discussion

By screening 111 whole genome sequences spanning the *Formica* phylogeny, we provide strong evidence that trans-species, haplotype-specific SNPs in the gene *knockout* that were originally identified and associated with colony social structure in five well-studied species (Brelsford et al. 2020) can be traced back to at least the common ancestor of extant *Formica* species (Fig. 1). By comparing differences between the genomes of individuals that are heterozygous and homozygous at *knockout* with a focus on the previously unsampled *neogagates-pallidefulva* clade (Borowiec et al. 2021), we discover additional clusters of trans-species, haplotype-specific SNPs that align to other regions of chromosome 3. This suggests first that all or part of the two *F. selysi* supergene variants formed in the common ancestor of all extant *Formica.* Second, these regions harbor additional genes that are candidates for contributing to the formation of the supergene and, perhaps, the contemporary phenotypes associated with it.

What conditions could lead to the maintenance of ancient trans-species, haplotype-specific SNPs distributed in several hotspots along chromosome 3 for approximately 30 Ma, through numerous speciation events? The patterns identified in this comparison are consistent with the predictions of the eroded strata model (predictions and empirical support are described in Table 1), but more information is needed to determine whether one or more inversions were already present in the common ancestor of contemporary *Formica* species. We emphasize the importance of investigating synteny in the genus to reconstruct the history of structural rearrangements between and within the Sp and Sm haplotype groups, because the peaks of differentiation spaced far apart on the *F. selysi* Sm reference genome could be in tight physical linkage on the Sp or Sm haplotypes in other *Formica* clades. The two largest clusters of trans-

species, haplotype-specific SNPs in the *neogagates-pallidefulva* clade occur near the ends of the *F. selysi* supergene, assuming that gene order is conserved between the *F. selysi* reference genome and members of this clade (Fig. 2). These could reflect the accumulation of mutations that are protected from rare recombination events due to their proximity to inversion breakpoints, particularly if the supergene initiated through one large inversion. However, we also find preliminary evidence that these regions have been duplicated in the putative Sp haplotype in several clades. In the haploid Sp males and Sp/Sp workers of *F. selysi*, *F. cinerea*, *F. lemani*, *F. exsecta*, and *F. truncorum*, many SNPs appear to be heterozygous in these two regions (Fig. 3). In the haploid Sp males especially, stretches of apparent heterozygosity likely reflect gene duplications. Here, Sp males appear to have two copies of each of these regions, one similar to the version present on the Sm haplotype and another that is substantially divergent. Gene duplications can accumulate in regions of suppressed recombination (e.g. Huu et al. 2020), as observed in the *S. invicta* supergene (Fontana et al. 2020). Future work on gene expression may reveal whether duplicated copies of these genes result in neofunctionalization, subfunctionalization or the formation of pseudogenes (e.g. Dang et al. 2019).

Through these analyses of additional *Formica* species, we have identified new genes that are highly differentiated between individuals that are Sm/Sm and Sm/Sp at *knockout*. These genes were not detected in the initial trans-species comparison reported by Brelsford et al. (2020). Two of these genes occur within the putative duplicated regions (*single-minded* and *FMRFaR*), while three more genes are found in the differentiated region upstream of *knockout* that appears to be duplicated on the Sp haplotype of *fusca* clade #3 (*AMGR10*, *tplus3b*, and an uncharacterized protein that may be a transcription factor). Like *knockout, single-minded* is associated with the development of the nervous system in *D. melanogaster* (Umetsu et al. 2006).

353 *FMRFaR* contributes to neurotransmitter release, and shapes behavioral responses to stimuli in
354 *D. melanogaster.* In particular, *FMRFaR* plays a role in muscle contraction and larval response
355 to light (Ravi et al. 2018; Klose et al. 2010). *AmGR10* is a gustatory receptor in *Apis mellifera*
356 and disrupting its activity affects the division of labor within honeybee hives (Paerhati et al.
357 2015). Specifically, this gene is expressed in nurses, and knockdown of gene expression causes
358 nurses to transition to foraging (Paerhati et al. 2015). Recent functional characterization of
359 *AmGR10* in *A. mellifera* reveals that this gene is a broadly-tuned amino acid receptor (Lim et al.
360 2019). This gene overlaps with the trans-species, haplotype-specific SNPs detected in our
361 comparisons (Fig. 3C), while two other genes are upstream and downstream of this area (Table
362 3). Further work is needed to examine the specific functions of these loci in *Formica* ants. One
363 major challenge in understanding the genetic underpinnings of colony-level traits, such as queen
364 number, is that these traits are difficult to assess in a standard gene knockdown assay.
365
366 *Additional patterns in the phylogenetic dataset*
367 Several additional patterns are visible in the comparison of *knockout* SNPs (Fig. 1). First,
368 in comparing the outgroup genotypes, we note that *Iberoformica* is homozygous for the reference
369 allele at 19 out of 20 sites, while *Polyergus* exhibits a mix of reference and alternate alleles.
370 While preliminary, this finding suggests that the Sm form of *knockout* may have been present in
371 the common ancestor of *Formica* and *Iberoformica.* This pattern was detected at the previously-
372 identified trans-species, haplotype-specific SNPs within *knockout,* but not in other parts of the
373 chromosome (Table S3). Second, samples from three of the 14 clades examined here contained
374 exclusively Sm/Sm homozygous individuals. We only sequenced three samples from the
375 *dakotensis* clade and four samples from *fusca* clade #6, but all 10 *difficilis* clade individuals were

homozygous for the *F. selysi* Sm/Sm reference alleles at *knockout*. Members of this clade are socially parasitic and include temporary social parasites and at least one permanent inquiline of other *Formica* species (Talbot 1976; Wilson 1976; Buschinger 1986; Borowiec et al. 2021). Previous research has suggested that polygyny can be a preadaptation for social parasitism (Alloway 1980; Elmes 1980; Bourke & Franks 1991). It's noteworthy that members of the *difficilis* clade are facultatively polygynous and evolved from a facultatively polygynous common ancestor (Borowiec et al. 2021) but that extant *difficilis* clade species may lack the supergene haplotype that is associated with polygyny in at least five congeneric species (Brelsford et al. 2020).

In the *neogagates-pallidefulva* clade, we detected a third haplotype at the *knockout* locus (Fig. 1). Given that these samples otherwise appear to be heterozygous at other trans-species, haplotype-specific SNPs detected within the *neogagates-pallidefulva* clade (Fig. 2A), we hypothesize that this haplotype reflects a clade-specific gene conversion or recombination event, and that most conserved, functionally relevant SNPs are within the 11,910,356-11,910,880 bp window. Finally, we note the dearth of samples that are apparently homozygous for the polygyne-associated haplotype at *knockout.* While Sp/Sp homozygotes are common in *F. selysi* (Purcell et al. 2014; Avril et al. 2019) and close relatives *F. cinerea* and *F. lemani* (Brelsford et al. 2020), our current sampling effort, which focused on maximizing species diversity, resulted in the addition of only one putative Sp/Sp homozygote in *fusca* clade #4. Samples were collected by several researchers without attention to colony social structure, so this pattern may simply reflect biased sampling. Alternatively, we speculate that the rarity of Sp/Sp in this sample raises the possibility that the Sp haplotype has experienced degeneration in some *Formica* lineages, similar to that observed in the *Solenopsis* supergene (Pracana et al. 2017; Stolle et al. 2019).

Limitations and future directions

Our study has several limitations that could influence the location and number of trans-species, haplotype-specific SNPs identified, but that are unlikely to affect the overall pattern that we described. We avoided assessing haplotype-specific variation on chromosome 3 for clades represented by few genomes because the small samples size could inflate the numbers of SNPs that show a false positive difference between alternative haplotypes. A subset of the trans-species, haplotype-specific SNPs identified within single clades (Figs. 2D-F), for example, are likely to be false positives. As we include more species spanning a deeper evolutionary history, our metric of assessing trans-species, haplotype-specific SNPs becomes highly conservative. In this analysis, we only include loci that show a perfect association with the *knockout* haplotype. Genotyping errors will inflate the number of false negatives, and this tendency will increase as the sample size grows. Finally, we aligned reads from all genomes to the *Formica selysi* reference genome (Brelsford et al. 2020), which is based upon an individual with the Sm/Sm genotype. While most sequences aligned to the reference genome, rapid divergence at some genomic regions could impede alignment for more distantly related species. All of our alignments assume the Sm haplotype orientation, which is collinear between *F. selysi* and *F. exsecta* (Brelsford et al. 2020). However, we know that there have been multiple rearrangements of chromosome 3 between Sm and Sp haplotypes in *F. selysi*. As a result, we cannot draw conclusions about the relative positions of the trans-species, haplotype-specific SNPs in all species or on both haplotypes. A useful complement to this study and future direction will be to examine structural variation on chromosome 3 across the genus *Formica.*

Conclusion

Overall, we demonstrate that genetic polymorphisms associated with variation in colony queen number in at least some *Formica* species (Purcell et al. 2014; Brelsford et al. 2020) likely predate the common ancestor of all extant *Formica*. Using a comparative approach, we have pinpointed additional genes within the supergene region that harbor trans-species, haplotype-specific SNPs in an early diverging, Nearctic *Formica* clade.

Acknowledgements: The authors thank members of the Brelsford and Purcell labs, including Z. Alam, E. Henderson, D. McGuire, N. Najar, M. Palanchon, D. Pierce, M. Sankovitz, G. Scarparo, and M. West, for helpful feedback on the manuscript. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. Research was supported by NSF-CAREER DEB-1942252 and U. S. Department of Agriculture National Institute of Food and Agriculture Hatch #CA-R-ENT-5126-H to JP, NSF GRFP DGE-1326120 to GLR, NSF DEB–1654829 and NSF CAREER DEB–1943626 to CR, and NSF DEB-1754834 to AB and JP.

Data Accessibility and Benefit-Sharing Statement: Sequences generated for this study will be placed on the NCBI SRA. Genotypes used in the analyses presented in this manuscript are provided in the supplementary materials files, in tables S2 and S3.

Formatted: Highlight

Author contributions: JP and AB conceived of the study. JP, MLB, CR, and AB collected and solicited the samples, MLB and CR identified the samples, and JP, MLB, and AB performed lab work to prepare samples for sequencing. AB and GLR analyzed the data. JP wrote the manuscript with advice from all authors.

References Cited

Alloway, T. M. (1980). The origins of slavery in Leptothoracine ants (Hymenoptera: Formicidae). *The American Naturalist*, *115*(2), 247–261.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389-3402.

Avril, A., Purcell, J., Béniguel, S., & Chapuisat, M. (2020). Maternal effect killing by a supergene controlling ant social organization. *Proceedings of the National Academy of Sciences*, *117*(29), 17130–17134. doi: 10.1073/pnas.2003282117

Avril, A., Purcell, J., Brelsford, A., & Chapuisat, M. (2019). Asymmetric assortative mating and queen polyandry are linked to a supergene controlling ant social organization. *Molecular Ecology*, *28*(6), 1428–1438. doi: 10.1111/mec.14793

461 Bargum, K., & Sundström, L. (2007). Multiple breeders, breeder shifts and inclusive fitness
462 returns in an ant. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1617), 1547–
463 1551. doi: 10.1098/rspb.2007.0295
464 Boc, A., Diallo, A. B., & Makarenkov, V. (2012). T-REX: A web server for inferring, validating
465 and visualizing phylogenetic trees and networks. *Nucleic Acids Research*, *40*(W1), W573–
466 W579. doi: 10.1093/nar/gks485
467 Borowiec, M. L., Cover, S. P., & Rabeling, C. (2021). The evolution of social parasitism in
468 *Formica* ants revealed by a global phylogeny. *Proceedings of the National Academy of*
469 *Sciences of the United States*, 118, e2026029118. doi: 10.1073/pnas.2026029118
470 Branco, S., Carpentier, F., Rodríguez de la Vega, R. C., Badouin, H., Snirc, A., Le Prieur, S., …
471 Giraud, T. (2018). Multiple convergent supergene evolution events in mating-type
472 chromosomes. *Nature Communications*, *9*(1), 2000. doi: 10.1038/s41467-018-04380-9
473 Brelsford, A., Purcell, J., Avril, A., Tran Van, P., Zhang, J., Brütsch, T., … Chapuisat, M.
474 (2020). An Ancient and Eroded Social Supergene Is Widespread across Formica Ants.
475 *Current Biology*, *30*(2), 304-311.e4. doi: 10.1016/j.cub.2019.11.032
476 Bourke, A. F., & Franks, N. R. (1991). Alternative adaptations, sympatric speciation and the
477 evolution of parasitic, inquiline ants. *Biological Journal of the Linnean Society*, *43*, 157–178.
478 Buschinger, A. (1986). Evolution of social parasitism in ants. *Trends in Ecology & Evolution*,
479 *1*(6), 155–160. doi: 10.1016/0169-5347(86)90044-3
480 Chapuisat, M., Bocherens, S., & Rosset, H. (2004). Variable Queen Number in Ant Colonies: No
481 Impact on Queen Turnover, Inbreeding, and Population Genetic Differentiation in the Ant
482 Formica Selysi. *Evolution*, *58*(5), 1064–1072. doi: 10.1111/j.0014-3820.2004.tb00440.x
483 Charlesworth, D. (2016). The status of supergenes in the 21st century: Recombination
484 suppression in Batesian mimicry and sex chromosomes and other complex adaptations.
485 *Evolutionary Applications*, *9*(1), 74–90. doi: 10.1111/eva.12291
486 Cohen, P., & Privman, E. (2020). The social supergene dates back to the speciation time of two
487 Solenopsis fire ant species. *Scientific Reports*, *10*(1), 11538. doi: 10.1038/s41598-020-
488 67999-z
489 Colecchia, D., Strambi, A., Sanzone, S., Iavarone, C., Rossi, M., Dall'Armi, C., … Chiariello,
490 M. (2012). MAPK15/ERK8 stimulates autophagy by interacting with LC3 and GABARAP
491 proteins. *Autophagy*, *8*(12), 1724–1740. doi: 10.4161/auto.21857
492 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Group, 1000
493 Genomes Project Analysis. (2011). The variant call format and VCFtools. *Bioinformatics*,
494 *27*(15), 2156–2158. doi: 10.1093/bioinformatics/btr330
495 Dang, V.-D., Cohanim, A. B., Fontana, S., Privman, E., & Wang, J. (2019). Has gene expression
496 neofunctionalization in the fire ant antennae contributed to queen discrimination behavior?
497 *Ecology and Evolution*, *9*(22), 12754–12766. doi: 10.1002/ece3.5748
498 DeHeer, C. J., & Herbers, J. M. (2004). Population genetics of the socially polymorphic ant
499 Formica podzolica. *Insectes Sociaux*, *51*(4), 309–316. doi: 10.1007/s00040-004-0745-1

Dhaygude, K., Nair, A., Johansson, H., Wurm, Y., & Sundström, L. (2019). The first draft genomes of the ant *Formica exsecta*, and its *Wolbachia* endosymbiont reveal extensive gene transfer from endosymbiont to host. *BMC Genomics* 20(1), 301. doi: 10.1186/s12864-019-5665-6

Elmes, G. W. (1980). Queen numbers in colonies of ants of the genus *Myrmica*. *Insectes Sociaux*, *27*(1), 43–60. doi: 10.1007/BF02224520

Fontana, S., Chang, N.-C., Chang, T., Lee, C.-C., Dang, V.-D., Wang, J. (2020). The fire ant social supergene is characterized by extensive gene and transposable element copy number variation. *Molecular Ecology, 29*(1), 105-120. doi: 10.1111/mec.15308

Guerrero, R.F., Rousset, F., & Kirkpatrick, M. (2012). Coalescent patterns for chromosomal inversions in divergent populations. *Philosophical Transactions of the Royal Society B, 367*, 430-438.

Hartmann, C., Landgraf, M., Bate, M., & Jäckle, H. (1997). Krüppel target gene knockout participates in the proper innervation of a specific set of Drosophila larval muscles. *The EMBO Journal*, *16*(17), 5299–5309. doi: 10.1093/emboj/16.17.5299

Hauschteck-Jungen, E., & Jungen, H. (1976). Ant chromosomes. I. The genus *Formica*. *Insectes Sociaux, 23*, 513-524.

Henderson, E., & Brelsford, A. (2021). Genomic differentiation across the speciation continuum in three hummingbird species pairs. *BMC Evolutionary Biology, 20*, 113. doi: 10.1186/s12862-020-01674-9

Hung, A.C.F. (1969). The chromosome numbers of six species of formicine ants. *Annals of the Entomological Society of America, 62*(2), 455-456.

Huu, C. N., Keller, B., Conti, E., Kappel, C., & Lenhard, M. (2020). Supergene evolution via stepwise duplications and neofunctionalization of a floral-organ identity gene. *Proceedings of the National Academy of Sciences*, *117*(37), 23148–23157. doi: 10.1073/pnas.2006296117

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., … ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*(7363), 203–206. doi: 10.1038/nature10341

Keller, L. (1993). The Assessment of Reproductive Success of Queens in Ants and Other Social Insects. *Oikos*, *67*(1), 177–180. doi: 10.2307/3545107

Klevernic, I. V., Martin, N. M. B., & Cohen, P. (2009). Regulation of the activity and expression of ERK8 by DNA damage. *FEBS Letters*, *583*(4), 680–684. doi: 10.1016/j.febslet.2009.01.011

Klose, M. K., Dason, J. S., Atwood, H. L., Boulianne, G. L., & Mercier, A. J. (2010). Peptide-Induced Modulation of Synaptic Transmission and Escape Response in Drosophila Requires Two G-Protein-Coupled Receptors. *Journal of Neuroscience*, *30*(44), 14724–14734. doi: 10.1523/JNEUROSCI.3612-10.2010

Kronforst, M. R., & Papa, R. (2015). The Functional Basis of Wing Patterning in *Heliconius* Butterflies: The Molecules Behind Mimicry. *Genetics*, *200*(1), 1–19. doi: 10.1534/genetics.114.172387

Küpper, C., Stocks, M., Risse, J. E., dos Remedios, N., Farrell, L. L., McRae, S. B., … Burke, T. (2016). A supergene determines highly divergent male reproductive morphs in the ruff. *Nature Genetics*, *48*(1), 79–83. doi: 10.1038/ng.3443

Lagunas-Robles, G., Purcell, J., & Brelsford, A. (2021). Linked supergenes underlie split sex ratio and social organization in an ant. *BioRxiv.* doi: 10.1101/2021.01.22.427864

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., … Andersson, L. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nature Genetics*, *48*(1), 84–88. doi: 10.1038/ng.3430

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. Retrieved from http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, J., Cocker, J. M., Wright, J., Webster, M. A., McMullan, M., Dyer, S., … Gilmartin, P. M. (2016). Genetic architecture and evolution of the S locus supergene in Primula vulgaris. *Nature Plants*, *2*(12), 16188. doi: 10.1038/nplants.2016.188

Lim, S., Jung, J., Yunusbaev, U., Ilyasov, R., & Kwon, H.W. (2019). Characterization and its implication of a novel taste receptor detecting nutrients in the honey bee, *Apis mellifera*. Scientific Reports, 9, 11620. 10.1038/s41598-019-46738-z

Paerhati, Y., Ishiguro, S., Ueda-Matsuo, R., Yang, P., Yamashita, T., Ito, K., … Suzuki, K. (2015). Expression of AmGR10 of the Gustatory Receptor Family in Honey Bee Is Correlated with Nursing Behavior. *PLOS ONE*, *10*(11), e0142917. doi: 10.1371/journal.pone.0142917

Pracana, R., Priyam, A., Levantis, I., Nichols, R. A., & Wurm, Y. (2017). The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. *Molecular Ecology*, *26*(11), 2864–2879. doi: 10.1111/mec.14054

Purcell, J., Brelsford, A., Wurm, Y., Perrin, N., & Chapuisat, M. (2014). Convergent Genetic Architecture Underlies Social Organization in Ants. *Current Biology*, *24*(22), 2728–2732. doi: 10.1016/j.cub.2014.09.071

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. doi: 10.1086/519795

Ramanauskas, K., & Igić, B. (2017). The evolutionary history of plant T2/S-type ribonucleases. *PeerJ*, *5*, e3790. doi: 10.7717/peerj.3790

Ramanauskas, K., & Igić, B. (in press, 2021). RNase-based self-incompatibility in cacti. *New Phytologist.* doi: https://doi.org/10.1111/nph.17541

Ravi, P., Trivedi, D., & Hasan, G. (2018). FMRFa receptor stimulated Ca2+ signals alter the activity of flight modulating central dopaminergic neurons in Drosophila melanogaster. *PLOS Genetics*, *14*(8), e1007459. doi: 10.1371/journal.pgen.1007459

580 Romiguier, J., Rolland, J., Morandin, C., & Keller, L. (2018). Phylogenomics of palearctic
581 Formica species suggests a single origin of temporary parasitism and gives insights to the
582 evolutionary pathway toward slave-making behaviour. *BMC Evolutionary Biology*, *18*(1), 40.
583 doi: 10.1186/s12862-018-1159-4
584 Rosengren, M., Rosengren, R., & Söderlund, V. (1980) Chromosome numbers in the genus
585 *Formica* with special reference to the taxonomical position of *Formica uralensis* Ruszk. and
586 *Formica truncorum* FABR. *Hereditas, 92*(2), 321-325.
587 Rosset, H., & Chapuisat, M. (2007). Alternative life-histories in a socially polymorphic ant.
588 *Evolutionary Ecology*, *21*(5), 577–588. doi: 10.1007/s10682-006-9139-3
589 Schwander, T., Libbrecht, R., & Keller, L. (2014). Supergenes and Complex Phenotypes.
590 *Current Biology*, *24*(7), R288–R294. doi: 10.1016/j.cub.2014.01.056
591 Seppä, P., Gyllenstrand, N., Corander, J., & Pamilo, P. (2004). Coexistence of the Social Types:
592 Genetic Population Structure in the: Ant Formica Exsecta. *Evolution*, *58*(11), 2462–2471.
593 doi: 10.1111/j.0014-3820.2004.tb00875.x
594 Stolle, E., Pracana, R., Howard, P., Paris, C. I., Brown, S. J., Castillo-Carrillo, C., … Wurm, Y.
595 (2019). Degenerative Expansion of a Young Supergene. *Molecular Biology and Evolution*,
596 *36*(3), 553–561. doi: 10.1093/molbev/msy236
597 Sun, S., Coelho, M. A., Heitman, J., & Nowrousian, M. (2019). Convergent evolution of linked
598 mating-type loci in basidiomycete fungi. *PLOS Genetics*, *15*(9), e1008365. doi:
599 10.1371/journal.pgen.1008365
600 Talbot, M. (1976). The Natural History of the Workerless Ant Parasite Formica Talbotae.
601 *Psyche: A Journal of Entomology*, *83*(3–4), 282–288. doi: 10.1155/1976/149241
602 Thompson, M. J., & Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity*,
603 *113*(1), 1–8. doi: 10.1038/hdy.2014.20
604 Tuttle, E. M., Bergland, A. O., Korody, M. L., Brewer, M. S., Newhouse, D. J., Minx, P., …
605 Balakrishnan, C. N. (2016). Divergence and Functional Degradation of a Sex Chromosome-
606 like Supergene. *Current Biology*, *26*(3), 344–350. doi: 10.1016/j.cub.2015.11.069
607 Umetsu, D., Murakami, S., Sato, M., & Tabata, T. (2006). The highly ordered assembly of
608 retinal axons and their synaptic partners is regulated by Hedgehog/Single-minded in the
609 Drosophila visual system. *Development*, *133*(5), 791–800. doi: 10.1242/dev.02253
610 Wang, J., Wurm, Y., Nipitwattanaphon, M., Riba-Grognuz, O., Huang, Y.-C., Shoemaker, D., &
611 Keller, L. (2013). A Y-like social chromosome causes alternative colony organization in fire
612 ants. *Nature*, *493*(7434), 664–668. doi: 10.1038/nature11832
613 Wilson, E.O. (1976). The first workerless parasite in the ant genus *Formica* (Hymenoptera:
614 Formicidae). *Psyche: A Journal of Entomology*, *83*(3-4), 277–281. doi: 10.1155/1976/149234
615 Yan, Z., Martin, S. H., Gotzek, D., Arsenault, S. V., Duchen, P., Helleu, Q., … Keller, L. (2020).
616 Evolution of a supergene that regulates a trans-species social polymorphism. *Nature Ecology*
617 *& Evolution*, *4*(2), 240–249. doi: 10.1038/s41559-019-1081-1

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. doi: 10.1093/bioinformatics/btt593

Figures and Tables

Fig. 1: The 22 SNPs with fixed differences between Sm and Sp haplotypes at the gene *knockout* that were conserved in 15 European species (Brelsford et al. 2020) are also conserved in North American *Formica* species. *Formica* species in our dataset cluster into 14 clades based on a neighbor joining tree. This tree aligns well with the best current *Formica* phylogeny, produced by Borowiec et al. (2021). Key divergence dates inferred by Borowiec et al. are indicated in grey circles on several nodes: node 1 is ~23-40 Ma, node 2 is ~20-30 Ma, and node 3 is ~17-27 Ma. Genotypes matching the reference genome (*F. selysi* Sm/Sm) are shown in green. Heterozygous loci are shown in light orange and loci homozygous for the alternate allele ('Sp/Sp') are shown in dark orange. White boxes represent missing data. Each clade is represented by a unique color bar (right), which we use consistently in other tables and figures to denote clade (Table S1, Fig. 2). Individual genomes that were previously analyzed by Brelsford et al. (2020) are shown in blue (monogyne origin), red (polygyne origin), and grey (social structure unknown) highlights over the species name.

Fig. 2: Comparing the five *knockout* heterozygotes and six *knockout* homozygotes from the *neogagates-pallidefulva* clade (A), we detect 429 clade-level trans-species, haplotype-specific SNPs. The majority of these are distributed from 2-12.5 Mbp along chromosome 3 with several notable peaks of divergence between alternative haplotypes at 2.3, 11.6, 11.9, and 12.4 Mbp. Species in this comparison share a common ancestor dating to about 24 Ma (Borowiec et al. 2021). We compare individuals with alternative *knockout* variants across the socially parasitic *Formica* lineages (B), which share a single common ancestor dating to about 18 Ma. We find peaks of divergence again at 2.3 and 11.6 Mbp, as well as many other clade-specific peaks. Comparing species from three *fusca* clades (C), we find only 68 trans-species, haplotype-specific SNPs. The majority of these SNPs are found at 2.3, 10.3, 11.6, and 11.9 Mbp (the latter includes *knockout*). These three clades share a common ancestor dating to about 17 Ma. We then investigate the distribution of trans-species, haplotype-specific SNPs in each of the three clades separately (D-F) to determine whether the patterns are consistent with the predictions of the 'eroded strata' model (Table 1). For *fusca* clade #3 (D, common ancestor dates to about 12 Ma), which includes *F. selysi*, we find clusters of trans-species, haplotype-specific SNPs spanning the 2-12.5 Mbp region containing the 'social' supergene (Purcell et al. 2014). We find a similar pattern in *fusca* clade #4 (E, common ancestor dates to about 13 Ma), except that we detect low levels of trans-species, haplotype-specific SNPs across the full extent of the chromosome. We find fewer SNPs in *fusca* clade #5 (F, common ancestor dates to about 14 Ma), with clear peaks of divergence at 2.3, 10.3, 11.6, and 11.9 Mbp. The colors of the data points reflect clade-specific colors (Fig. 1, Table S1).

Fig. 3: Trans-species, haplotype-specific SNPs that are present across four well-sampled clades, including *neogagates-pallidefulva, exsecta*, *rufa,* and *fusca* clade #3 are found in several clusters along chromosome 3 (A). In the latter three clades, we are comparing Sp males and Sp/Sp workers with Sm males and Sm/Sm workers. Some peaks of divergence, most notably the

region at 2.3 Mbp, show an unusual pattern of apparent heterozygosity in the Sp haploids, suggesting possible gene duplications in these regions (circles). We also detected some peaks of divergence that appear to include gene duplications in only one clade, *fusca* clade #3 (squares). There are relatively few trans-species, haplotype-specific SNPs that never appear heterozygous in the Sp males (triangles). For the regions with the greatest number of trans-species, haplotype-specific SNPs, we show zoomed in views of the SNPs near the genes *single-minded* (B) and *AmGR10* (C). Exons in these genes are shown as gray bars (B and C). The arrow below each figure indicates the orientation of the gene. We indicate the positions of the seven genes identified in Table 3 (A).

673 **Table 1:** Predictions of and empirical evidence consistent with the eroded strata model at three distinct time points. Visual predictions
674 of this verbal model and additional background are presented by Brelsford et al. (2020, Figure 4).

Comparison	Expected pattern	Empirical evidence consistent with the prediction	Additional expectations
Within polymorphic species	Suppressed recombination results in differentiation between alternative haplotypes across the length of the supergene.	F_{ST} between individuals from single- and multiple-queen colonies is elevated from 2-~12.5 Mbp along chromosome 3 in five species (Brelsford et al. 2020, Figure 1). Throughout this 10.5 Mbp supergene region, linkage disequilibrium is high in *F. selysi* (Figure 2B in Purcell et al. 2014).	Recent recombination or gene conversion events may be detected as a polymorphism within one haplotype. An example was observed in the *F. selysi* Sp haplotype (Purcell et al. 2014, Figure S1J).
Between closely related polymorphic species	Comparing alternative trans-species haplotypes will reveal many SNPs that are shared across species. There may be some areas within the region of suppressed recombination that lack such SNPs.	The related species *F. selysi*, *F. cinerea*, and *F. lemani* share trans-species, haplotype-specific SNPs along long stretches of the supergene (Brelsford et al. 2020, Figure 3). Within relatively young clades (fusca clades #3, 4, and 5), we similarly see trans-species, haplotype-specific SNPs spanning most of chromosome 3 (this study, Figures 2D-F).	Clades with a longer evolutionary history or more representatives in the analysis will have smaller regions harboring trans-species, haplotype-specific SNPs within the region of suppressed recombination.
Between more distantly related groups of polymorphic species.	As the trans-species analysis is expanded to include groups sharing an increasingly ancient common ancestor, the comparison will encompass a larger number of rare, branch-specific recombination events if we assume that rare recombination occurs at a slow and steady rate after the inversion forms. These regions will not harbor trans-species, haplotype-specific SNPs. Such SNPs will only be found in parts of the chromosome that have remained differentiated between the alternative haplotypes in all branches of the tree.	When comparing across the genomes of 15 European *Formica* species, Brelsford et al. (2020) found 22 trans-species, haplotype-specific SNPs in the gene *knockout* (Brelsford et al. 2020, Figure 1). Here, we greatly expand trans-species analyses of the SNPs in *knockout* (this study, Figure 1), as well as SNPs across chromosome 3 in the *neogagates-pallidefulva* clade (Figure 2A), the socially parasitic clade (Figure 2B), and across *fusca* clades #3-5 (Figure 2C).	At this scale, we expect the trans-species, haplotype-specific SNPs to either be localized in parts of the chromosome that are under selection or in areas that cannot recombine due to structural constraints (e.g. inversion breakpoints). We note that there may also be differing structural rearrangements between alternative haplotypes in different clades.

Table 2: Genome sequencing libraries were prepared in three groups, as shown here. Details about each sample are shown in table S1.

Number of genomes	Source	Library preparation method	Depth (range)
21	Brelsford et al. 2020	TruSeq	7-18.2x
15	collected by Purcell and Brelsford	TruSeq	5.3-15x
1	collected by Purcell and Brelsford	Nextera	8.7x
80	collected by Borowiec, Cover and Rabeling	Kapa	1.5-30.7x

Table 3: Regions of the supergene that harbor large numbers of trans-species, haplotype-specific SNPs overlap with several genes that could play a functional role in the evolutionary origin of the supergene or in the contemporary maintenance of polymorphism. We briefly summarize the functions of these genes in model systems when possible.

gene	ortholog name	trans-species SNP position(s)	putative function	notes	citation
Knockout (Drosophila melanogaster)	STOX1 (*Homo sapiens*)	Scaffold03: 11,910,116-11,911,137	motor neuron development	conserved trans-species fixed SNPs detected in 11/14 *Formica* clades	Brelsford et al. 2020; Hartmann et al. 1997
Single-minded (*D. melanogaster*)	SIM1 (*H. sapiens*)	Scaffold03: 2,261,463-2,275,100	neurogenesis, transcription	Sp version appears to contain a duplication (Fig. 3)	this article; Umetsu et al. 2006
AmGR10 (*Apis mellifera*)	NA	Scaffold03: 11,647,776-11,683,774	gustatory receptor, implicated in division of labor	*F. exsecta* annotation may have incorrectly split this gene in two	this article; Paerhati et al. 2015; Lim et al. 2019
tplus3b (*D. melanogaster*)	*mapk15* (*H. sapiens*)	Scaffold03: 11,647,776-11,683,774	ATP binding, DNA and telomere repair, regulation of autophagy	function examined in humans, only the last intron overlaps with conserved trans-species fixed SNPs	this article; Colecchia et al. 2012, Klevernic et al. 2009
zinc finger protein 148-like	NA	Scaffold03: 11,647,776-11,683,774	transcription regulation	just upstream of putative AmGR10	this article
FMRFaR (*D. melanogaster*)	NA	Scaffold03: 12,353,285-12,369,100	neuropeptide receptor, locomotion behavior, larval response to light	part of this region may have been duplicated on Sp haplotype in some clades (Fig. 3)	this article; Ravi et al. 2018; Klose et al. 2010
Uncharacterized protein (LOC115241360)	NA	Scaffold03: 10,262,710-10,263,020	possible transcription factor		this article