Genomic architecture and evolutionary dynamics of a social niche polymorphism in the California harvester ant, *Pogonomyrmex californicus*

Running title:

California harvester ant population genomics

Mohammed Errbii[1], Ulrich R. Ernst[1,§], Aparna Lajmi[2], Jürgen Gadau[1,3,‡], Lukas Schrader[1,3,‡]

[1] Molecular Evolution and Sociobiology Group, Institute for Evolution & Biodiversity, University of Münster, Hüfferstr. 1, DE-48149 Münster, Germany

[2] Department of Evolutionary and Environmental Biology, Institute of Evolution, University of Haifa, Haifa, Israel

[3] lukas.schrader@uni-muenster.de; gadauj@uni-muenster.de

[‡] equal contribution

[§] current address: Apicultural State Institute, University of Hohenheim, August-von-Hartmann-Str. 13, DE-70599 Stuttgart, Germany

Abstract

The societies of social insects are highly variable, including variation in the number of reproductives in a colony. In the California harvester ant, *Pogonomyrmex californicus* (Buckley 1867), colonies are commonly founded by a single queen (haplometrosis, primary monogyny). However, in some populations in California (USA), two or more queens cooperate in colony founding (pleometrosis) and continue to share a nest over several years (primary polygyny). Here, we use population genomics and linkage mapping to study the evolutionary dynamics and genetic architecture of this social niche polymorphism. Our analyses show that both populations underwent consecutive bottlenecks over the last 100,000 generations, particularly decreasing population size in the P-population and that the two populations diverged until 1,000 generations ago, after which gene flow increased again and we found signs of recent genetic admixture between the two populations. We further uncover an 8 Mb non-recombining region segregating with the observed social niche polymorphism, showing characteristics of a supergene comparable to the ones underlying social niche polymorphism in other ant species. In addition, 57 genes in five genomic regions outside the supergene show signatures of a selective sweep in the P-population, some of which are differentially expressed between haplo- and pleometrotic queens during colony founding. Our findings expose the social niche polymorphism in *P. californicus* as a polygenic trait involving a supergene.

Keywords: haplometrosis, pleometrosis, supergene, social niche polymorphism, population genomics, monogyny, polygyny, *Pogonomyrmex*

Introduction

An important component of the social organization of eusocial species is the number of reproductive individuals (Wilson 1971; Hölldobler and Wilson 1977; Boomsma et al. 2014). This number is usually described as fixed for a given species with insignificant variations. However, with the advent of genetic marker systems it became clear that many species show a range of social organization related to queen number both in terms of their species range (e.g. *Formica exsecta* (Seppä et al. 2004), *Solenopsis invicta* (Ross et al. 2007), *Messor pergandei* (Helms et al. 2013), *Pogonomyrmex californicus* (Overson et al. 2016)) and colony stage (secondary polygyny or monogyny) (Boomsma et al. 2014). This variation can be considered as specializations to different social niches, i.e. "the set of social environments in which the focal individual has non-zero inclusive fitness" (Saltz et al. 2016). For example, if queens become tolerant towards co-founding queen(s), a novel social niche can be established (pleometrosis = multiple queens found a colony together) that is distinct from the original niche (haplometrosis = queens found a colony alone), where queens kill potential co-foundresses either immediately or after the first workers emerge (Bernasconi and Strassmann 1999; Overson et al. 2014; Overson et al. 2016). Species that show intraspecific social niche variation are ideal to study the genetic architecture and evolutionary dynamics underlying such transitions. For example, recent population genomic studies in *Solenopsis* fire ants (Ross et al. 2007; Yan et al. 2020) and *Formica* wood ants (Purcell et al. 2014; Brelsford et al. 2020) have provided significant insights into the origin, evolution, and geographic distribution of social niche variation (monogyny vs polygyny) in ants.

The California harvester ant (*Pogonomyrmex californicus* (Buckley 1867)) is a widespread harvester ant that occurs throughout the Southwestern USA and Northwestern Mexico (Johnson 2002) and varies in terms of monogyny and polygyny. However, in contrast to *Solenopsis invicta* and *Formica* spp. this variation is based on differences in colony founding behavior, i.e. haplo- vs pleometrosis (Rissing et al. 2000; Johnson 2004; Overson et al. 2016; Haney and Fewell 2018), and not re-adoption of queens or colony budding (Glancey and Lofgren 1988; Rosengren et al. 1993). Queens of polygynous colonies and pleometrotic foundress associations in *P. californicus* are usually unrelated (Overson et al. 2016). Hence, both represent multifamily groups where unrelated queens and workers, cooperate (Bernasconi and Strassmann 1999). Haney and Fewell (2018) found that higher colony densities and lower resource availabilities, i.e. harsher conditions, correlate with polygyny, suggesting that haplo- and pleometrosis are adaptations to divergent habitats. Haplometrosis and monogyny is the norm in *P. californicus* (Johnson

2004). In Southern California two geographically isolated populations, that are almost fixed for haplometrosis (primary monogyny) or pleometrosis (primary polygyny) occur in close proximity (app. 50 km; Overson et al., 2014; Rissing et al., 2000). Whether both social niches can co-exist in a population, depends on ecological factors (e.g. nest site limitation, population density, resource limitations (Herbers 1986; Haney and Fewell 2018)), but could also depend on so far unidentified genetic factors (supergenes, genotype specific mortality, (Purcell et al. 2014; Schwander et al. 2014). It has been shown that supergenes, i.e. genomic regions containing clusters of linked loci (Schwander et al. 2014) play an important role in the evolution and maintenance of complex trait polymorphisms (e.g. Hohenlohe et al. 2010; Nadeau et al. 2012; Brelsford et al. 2020; Matschiner et al. 2021), and in fact, the genetic foundations of social niche polymorphisms in ants studied so far are supergenes (Wang et al. 2013; Purcell et al. 2014; Brelsford et al. 2020; Yan et al. 2020).

In this study, we explored the population dynamics, evolutionary history, and genetic architecture of the haplometrosis/pleometrosis social niche polymorphism in *P. californicus*. Coalescent modeling shows divergence between the two populations and their associated social niche phenotypes began 100,000 generations ago, but that gene flow between both populations has increased over the last 1,000 generations. We find evidence for recurring genetic bottlenecks in both populations, most likely reflecting the preference of *P. californicus* for colonizing disturbed habitats. Based on extended haplotype analyses and linkage mapping, we identify a non-recombining region of app. 8 Mb that segregates with our observed social niche polymorphism. This region shows all characteristics of a supergene, comparable to the ones underlying social niche polymorphism in other ants. However, unlike in *Solenopsis* and *Formica* we also find signatures of selective sweeps in the pleometrotic population beyond the non-recombining region, suggesting that social niche polymorphism in *P. californicus* is a polygenic trait involving a supergene.

87 **Material and Methods**

88 *Ant Collection*

89 We collected several hundred gynes of *Pogonomyrmex californicus* after their nuptial flights in 2018 and
90 2019 at Pine Valley (32.822 N, -116.528 W; predominantly pleometrotic population, P-population) and
91 Lake Henshaw resort (33.232 N, -116.760 W; predominantly haplometrotic population, H-population),
92 California, USA, which are 50 km apart. Ants were brought to laboratories at Arizona State University,
93 Tempe, Arizona, USA and housed alone or in pairs in observation nests. Observation nests were so called
94 sandwich nests that consist of a sand layer between two glass plates sealed on all four sides with plastic
95 spacers. After one week of behavioral observation, queens were snap-frozen in liquid nitrogen and
96 transferred to vials containing 0.5 ml RNAlater (Invitrogen). There are no legal regulations concerning the
97 collection or treatment of ants in the USA or Germany.

98 To test whether individual queens indeed behaved as predicted (queens from haplometrotic populations
99 are expected to be aggressive when paired with any other queen, and queens from pleometrotic
100 populations are expected to cooperate when forced into one nest with another queen), we observed the
101 behavior of paired queens using scan sampling (for more details see Clark and Fewell 2014; Overson et al.
102 2014; Helmkampf et al. 2016).

103

104 *Barcoding*

105 All ants were barcoded using partial *Cytochrome Oxidase I* (COI) to confirm species identification and
106 determine mitochondrial variation in both populations. In brief, we extracted DNA from a single leg
107 following a Chelex protocol (Gadau 2009a), performed a PCR and sequenced the PCR products using
108 Sanger sequencing. We mixed 1 µl of DNA extract, 2 µl Colorless GoTaq Reaction Buffer (7.5 mM $MgCl_2$;
109 Promega, Walldorf, Germany), 1 µl 25 mM $MgCl_2$, 0.6 µl 1.25 mM dNTPs, 0.2 µl of each primer (20 µM,
110 HCO: TAA ACT TCA GGG TGA CCA AAA AAT CA; LCO: GGT CAA CAA ATCATA AAG ATA TTG G), 0.06 µl GoTaq
111 G2 DNA Polymerase Buffer (5 u/µl; Promega, Walldorf, Germany) and 4.94 µl demineralized H_2O. The PCR
112 program consisted of an initial 3 min melting at 94 °C; 38 cycles of 1 min at 94 °C, 1 min at 45 °C, and 1.5
113 min at 72 °C; and a final extension step for 5 min at 72 °C. We removed remaining primers and
114 dephosphorylated single nucleotides by incubating 5 µl PCR product with 0.25 µl Exonuclease 1 (5 u;
115 Thermofisher, Dreieich, Germany) and 0.5 µl Shrimp alkaline phosphatase (5 u; Thermofisher, Dreieich,

116 Germany) for 15 min at 37 °C before deactivating the enzymes for 15 min at 85 °C. We added 5 µl of primer
117 HCO or LCO, sequencing the PCR product in both directions at the European Custom Sequencing Centre,
118 Cologne, Germany. Sequences of partial mitochondrial *cytochrome oxidase I* (COI) were blasted online at
119 NCBI against the non-redundant nucleotide collection (nt) using blastn with standard settings for short
120 sequences (Altschul et al. 1990). All sequences were aligned and based on this alignment a neighbor joining
121 tree was calculated using the online tool of MAFFT v. 7 (Katoh et al. 2018).

122

123 *DNA extraction, library prep and genome sequencing*

124 DNA was extracted using a phenol-chloroform-extraction protocol following Gadau (2009a). Libraries were
125 constructed using NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina (NEB #E7805; New England
126 BioLabs, Ipswich, USA) according to manufacturer's protocol. DNA extracts and libraries were controlled
127 for purity and size using NanoPhotometer Pearl (Implen, Munich, Germany) and Bioanalyzer 2100 (Agilent,
128 Santa Clara, USA). Libraries were paired-end sequenced (2x 75 bp) at the Genomic Core Facility of the
129 University of Münster (Germany), using a NextSeq 500 system with NextSeq 500 High Output Kit v2.5
130 (Illumina, San Diego, USA). We sequenced 16 queens from a haplometrotic population and 19 from a
131 pleometrotic population (total n = 35) with an average coverage of more than 12x (Supporting Table S1).

132

133 *Genome assembly and annotation*

134 We sequenced and assembled the genome of *P. californicus* ("Pcal2") using data from MinION long reads
135 (MinION, Oxford Nanopore Technologies, Oxford, UK), Illumina short reads, and previously published 10X
136 Chromium linked reads (Bohn et al. 2021). For long read library preparation, we sampled three white
137 worker pupae from a pleometrotic colony ("Pcal18-02") in liquid nitrogen for high-molecular weight DNA
138 extraction using a phase-separation protocol, optimized for DNA extraction from ants. After quality control
139 with agarose gels, Nanodrop and Qubit measurements, we used 2 µg of DNA for library preparation with
140 the Ligation Sequencing Kit (Oxford Nanopore) SQK-LSK109. The resulting library was sequenced for 48 h
141 on a MinION sequencer, using a single flow cell. After basecalling with guppy (v. 4.0.14) we retrieved
142 1,650,688 reads (mean read length 4.9 kb) with a mean read quality of 13.4. Read length N50 was 18.5 kb
143 and a total of 8.1 Gb of data was generated. We used PoreChop (v. 0.2.4) to remove adapter

contaminations from MinION reads before further filtering the data with FiltLong (v. 0.2.0). FiltLong allows selection of high quality reads based on comparison with short-read data of the same species. We sampled paired-end Illumina data of *P. californicus* queens for FiltLong from three samples (P50Y, P67P, P55R). We assembled the genome from the ~30x Nanopore data using the following strategy. We created one assembly using CANU (v. 2.0) and a second assembly using wtdbg2 (v. 2.5). Both assemblies were processed and optimized as follows: After contig assembly, we used SSPACE for scaffolding, LR_closer for gap filling, and ntedit for polishing with long-read data. We then used short-read, short-insert Illumina data from three samples (P50Y, P67P, P55R) for scaffolding with SSPACE before three rounds of polishing with Pilon. The polished assembly was super-scaffolded with publicly available 10X Chromium data for *P. californicus* (accession SRX8051306, Bohn et al. 2021) using scaff10x and subsequent correction with break10x. We used an in-house, blastn-based pipeline to detect and remove bacterial contaminations from the assemblies. Finally, we combined the CANU-based and the wtdbg-based assemblies with ntjoin, using the CANU-based assembly as reference. The joint assembly was filtered for redundancy using funannotate clean (v. 1.7.1), removing four scaffolds.

Genome annotation

We annotated repeats and transposable elements in the genome using a combination of de-novo, structural and homology-based approaches. We ran RepeatModeler2 (v. 2.0.1) and EDTA (v. 1.8.3) on the genome assembly to generate de-novo repeat libraries. Libraries from both tools were combined and redundancies removed with CD-HIT (v. 4.8.1). We classified repeats using a combination of PasteClassifier (included in REPET v. 2.5), DeepTE (github.com/LiLabAtVT/DeepTE) and TEsorter (v. 1.2.5.2). We removed putative host proteins from the library by blasting all unclassified repeats against uniprot_arthropoda (uniprot-arthropoda+6656 downloaded 27-Apr-2018) with “blastx -soft_masking false -evalue 1e-10”. We combined our de-novo library with public resources from RepBase (RepBase25.04.invrep.fa) and Dfam (Dfam3.0.arthropod.fa). Finally, using this combined library, we annotated the genome with RepeatMasker (v.open-4.0.7) with -s -gff -a -inv -small -excln -cutoff 250 -nolow. The produced annotation was subsequently improved further with One-code-to-find-them-all (v. 1.0).

Protein-coding and tRNA genes were annotated with funannotate v. 1.7.1 in the soft-masked genome. We used short-read Illumina paired-end RNAseq data of 18 heads of *P. californicus* (Ernst et al., unpublished)

173 and previously published long-read minION RNAseq data of a pool of workers (Bohn et al. 2021) for
174 training. Homology-based gene predictions were included by running GeMoMa v. 1.7 on a set of five
175 published ant genomes (*Pogonomyrmex barbatus*, GCF_000187915.1; *Ooceraea biroi*, GCF_003672135.1;
176 *Camponotus floridanus*, GCF_003227725; *Solenopsis invicta*, GCF_000188075.2, *Acromyrmex echinatior*,
177 GCF_000204515.1). Funannotate initially produced 26,408 gene models. After functional annotation with
178 interproscan, EggNog, and funannotate, we removed any protein-coding gene that did not have any
179 functional annotation or known ortholog and did not have any RNAseq coverage. Of the predicted genes,
180 9,435 had at least some RNAseq support and 15,110 showed at least some homology to known proteins
181 (with E-value < 1E-5 against UniRef90 Release: 2020_04, 12-Aug-2020). 1,279 genes had RNAseq support
182 but were not homologous to any protein in UniRef90.

183

184 *Mapping, SNP calling and filtering*

185 We used FastQC (v. 0.11.7) to inspect the quality of the raw short read data. Short and low quality reads
186 were removed using Trimmomatic (v. 0.38) (parameters: *SLIDINGWINDOW*:4:20 *MINLEN*:40) to only keep
187 paired and high quality reads (Bolger et al. 2014). The resulting paired reads were mapped to the newly
188 produced assembly of *P. californicus* (Pcal2) using BWA-MEM (v. 0.7.17-r1188) with default parameters (Li
189 and Durbin 2009). We assessed the quality of the resulting alignments using QualiMap (v. 2.2.1)
190 (Okonechnikov et al. 2016).

191 After marking duplicates using Picard's MarkDuplicates (v. 2.20.0), we performed joint-variant calling
192 across all samples, in two rounds.

193 We first generated a high quality set of SNPs that we used for base quality recalibration in the second pass.
194 In each pass, we used GATK's HaplotypeCaller (v. 4.1.2.0) with default parameters (McKenna et al. 2010;
195 Van der Auwera et al. 2013). Using this approach, we recovered a total of 1,486,819 SNPs that we hard-
196 filtered using the GATK VariantFiltration tool with the following parameters: QD < 2.0 || FS > 60.0 ||
197 MQRankSum < -5.0|| ReadPosRankSum < -5.0 || SOR > 3.0 || MQ < 40.0. We based our choice of these
198 parameters on a visual inspection of the distribution of annotation values of the raw SNPs as
199 recommended by the GATK documentation (Depristo et al. 2011).

200 The 1,359,080 SNPs that have passed the applied hard filters were further filtered using VCFtools (v.
201 0.1.16) (Danecek et al. 2011) and we kept only biallelic SNPs occurring in at least 85% of samples, with a

202 minor allele count (MAC) ≥ 3 and a site depth of 3X to 24X, resulting in 646,060 SNPs. These cut-offs were
203 chosen based on the distribution of these parameters in the raw SNPs left after applying the hard filters.
204 After a visual inspection of mapping quality across the genome (Supporting Figure S1), we excluded short
205 scaffolds with low mapping quality and considered 613,411 SNPs identified in the 25 largest scaffolds
206 (representing 222.6 Mb, i.e. 88.2% of the assembly) for further analyses.

207

208 *Statistical phasing*

209 As fineSTRUCTURE and the Multiple Sequential Markovian Coalescence modeling (MSMC2)
210 require phased haplotype data of single chromosomes as input, we used SHAPEIT (v. 2.r904) (Delaneau et
211 al. 2012) to statistically phase the 25 largest scaffolds of the *P. californicus* genome. Since a reference
212 panel is not available for *P. californicus*, we used the 35 single individual genomes that we sequenced in
213 this study for phasing. To produce a reliable set of phased genomes, we only considered 210,013 SNPs
214 with no missing data (i.e. genotyped in all 35 queens) for phasing. We separated the SNP dataset
215 genotyped across all 35 individuals into scaffolds using BCFtools (v. 1.9) and performed phasing by running
216 SHAPEIT (-*check*, -*phase* and -*convert*) on each scaffold separately.

217 Scaffolds 22 and 23 displayed an atypical architecture potentially representing a supergene (discussed
218 further below). Therefore scaffolds 22 and 23 as well as three other small scaffolds were excluded and we
219 only used phased SNPs contained in the 20 largest scaffolds for fineSTRUCTURE and MSMC2 analyses.

220

221 *Population structure analyses*

222 We used PLINK to LD-prune the 613,411 SNPs at a threshold of $r^2 < 0.2$, using the --*indep-pairwise* 1 kb 1
223 0.2 command. This operation yielded 314,756 LD-pruned SNPs that we then used to perform a Principal
224 Component Analysis (PCA) using the --*pca* command of PLINK (v. 1.90p) (Chang et al., 2015; Purcell et al.,
225 2007) and only considered 4 eigenvectors. The LD-pruned SNP dataset was also used to run ADMIXTURE
226 (v. 1.3.0) (Alexander et al. 2009) considering values of K from 2 to 4, and the best result was chosen based
227 on the lowest value for the cross-validation error.

228 To better characterize the relationships among queens in our sample, we used the haplotype-based
229 approach implemented in fineSTRUCTURE (Lawson et al. 2012). In this method, each individual is

230 considered in turn as a recipient, whose chromosomes are reconstructed as a combination of haplotypes
231 from all other individuals (the donors), using ChromoPainter (Lawson et al. 2012). Closely related
232 individuals are expected to share more haplotype blocks with each another. The final output is a
233 coancestry matrix that can be visualized as a heat map summarizing the number of shared haplotypes
234 between all pairs of individuals in the analyzed data sets.

235 We first converted SHAPEIT output to ChromoPainter's PHASE format using the *impute2chromopainter.pl*
236 provided by the developers of fineSTRUCTURE. We ran fineSTRUCTURE (v. 4.1.0) under the 'linked' mode
237 with default parameters. In our model, we assumed a uniform recombination rate, based on the estimate
238 of 14 cM/Mb found in *Pogonomyrmex rugosus* (Sirviö et al. 2011). The coancestry matrix plot was
239 produced in R (v. 4.0.2) (R Core Team 2020) using an R script, also provided with fineSTRUCTURE
240 [https://people.maths.bris.ac.uk/~madjl/finestructure/finestructureR.html; last accessed on March 11,
241 2021].

242

243 *Demographic population history analysis*

244 We inferred demographic population history from individual genomes using the Multiple Sequential
245 Markovian Coalescence modeling (MSMC2) (Schiffels and Durbin 2014). We used phased data generated
246 using SHAPEIT from four single individual queens (i.e. eight phased haplotypes) from the haplometrotic
247 (H123G, H118W, H73W and H123W; mean coverage 18X) and pleometrotic populations (P22P, P134P,
248 P117P and P109R; mean coverage ≥ 15X).

249 Mappability masks for each scaffold were generated using SNPABLE
250 [http://lh3lh3.users.sourceforge.net/snpable.shtml; last accessed on March 11, 2021] and following the
251 steps described by its developer.

252 Effective population size (N_e) and the relative cross-coalescence rate (rCCR) were then calculated using
253 MSMC2 following the developer`s guidelines (Schiffels and Durbin 2014; Schiffels and Wang 2020). We ran
254 1,000 bootstrap replicates, each time randomly sampling 60 Mb for each of the analyzed genomes. To
255 obtain the time in generations, we assumed a similar mutation rate as in honey bees (Yang et al. 2015),
256 i.e. $3.6 * 10^{-9}$ mutations * nucleotide^{-1} * generation^{-1}.

257

258 *Population genomic estimates*

259 We used VCFtools (v. 0.1.16) (Danecek et al. 2011) to estimate nucleotide diversity (π) (*--window-pi)* and
260 Tajima's *D (--TajimaD*) within each population using 100 kb non-overlapping windows across the genome.
261 VCFtools' *--het* function was employed to calculate heterozygosity and inbreeding coefficient on a per
262 individual basis. To reduce the effects of mapping errors, windows where the average mapping quality is
263 below 40 and/or show extreme depth of coverage and/or SNP number were excluded from further
264 analyses (Supporting Figure S2).

265 Linkage disequilibrium (LD) decay was estimated by calculating pairwise r^2 values between all markers
266 within 200 kb windows using VCFtools' *--hap-r2* function in each of the two populations. We used SNP
267 markers with a minor allele frequency (MAF) of 0.2 to minimize the effects of rare variants. We next
268 averaged r^2 values between all pairs of markers in 100 bp distance bins for ease of visualization.

269 We tested for statistical differences between populations by performing a Wilcoxon rank sum test with
270 "Bonferroni" adjustments of the p-values in R (v. 4.0.2) (R Core Team 2020).

271

272 *Genome scan for selection*

273 To identify candidate genes potentially responsible for the social polymorphism in *P. californicus*, we
274 screened genomes of both populations for evidence of selection using an F_{st} outlier approach and by
275 identifying regions of extended haplotype homozygosity (selective sweeps). We screened the genome for
276 loci showing signatures of selection using an F_{st} outlier approach comparing the haplometrotic and
277 pleometrotic populations (Kelley et al. 2006; Akey et al. 2010; Kolaczkowski et al. 2011; Fabian et al. 2012).

278 Prior to computing F_{st} values between the two populations, we estimated relatedness between all pairs of
279 queens using PLINK's *--make-rel* function (Chang et al., 2015; Purcell et al., 2007). Based on visual
280 inspection of the distribution of relatedness values, we excluded queens from pairs with relatedness > 0.4
281 leaving us with 17 (of 19) and twelve (of 16) queens from the pleometrotic and haplometrotic populations,
282 respectively. Moreover, we excluded one queen from the pleometrotic population as our population
283 structure analysis indicated that it was an F1 hybrid (P99P).

284 We calculated mean F_{st} in 100 kb non-overlapping windows across the genome using VCFtools' functions
285 *--weir-fst-pop* and *--fst-window-size* (v. 0.1.16) (Danecek et al. 2011). From the empirical distribution of F_{st}
286 values, we considered the top 5 % of these windows as differentiated (Figure 4A).

287 To determine in which population the differentiated windows are under selection, we computed
288 nucleotide diversity π and Tajima's *D* in 100 kb non-overlapping windows across the genome and reasoned
289 that these regions should show a reduced Tajima's *D* (typically negative values) and nucleotide diversity in
290 the population where they are subject of selection. Genomic windows failing these conditions were
291 flagged as ambiguous and were discarded from further processing, resulting in a total of 51 candidate
292 regions for evolving under divergent selection between the populations (Supporting Table S2).

293 To detect signatures of selective sweeps, we further applied the cross-population extended haplotype
294 homozygosity (*xpEHH*) and the site-specific extended haplotype homozygosity (*EHHS*) tests (Sabeti et al.
295 2007; Tang et al. 2007). The *xpEHH* compares the length of haplotypes between populations to detect
296 selective sweeps in which the selected allele has approached or reached fixation in at least one population
297 (Sabeti et al. 2007). The *EHHS* on the other hand is defined as the decay of identity of haplotypes
298 surrounding a core marker of a population as a function of distance (Tang et al. 2007).

299 The *xpEHH* analyses were performed on the phased data from unrelated queens using the R package rehh
300 (v. 3.1.2) (Gautier et al. 2017) with the following settings: *polarize_vcf=FALSE*, *min_maf=0.05*,
301 *scalegap=20,000* and *maxgap=200,000*. The latter two options were used in order to reduce false positives
302 caused by large gaps between markers (Voight et al. 2006). We used the Benjamani & Hochberg method
303 for FDR correction and set a p-value of 0.05 as a threshold to identify outlier SNPs. Outliers occurring within
304 100 kb of each other were clustered into one outlier region under selection, resulting in 32 clusters
305 distributed across multiple scaffolds (Supporting Table S3). Regions that are within -/+ 100 kb around
306 the most significant SNP of each cluster were kept as candidates for positive selective sweeps.

307 We used BEDtools' *intersect* function (v. 2.29.2) (Quinlan and Hall 2010) to extract regions that overlap
308 between the two selection screen approaches, producing 17 candidate regions (Supporting Table S4). We
309 excluded an ambiguous region on scaffold 11 (position: 1,200,000-1,349,402) that showed opposing
310 pattern of selection (i.e. F_{st} suggests selection in the pleometrotic population, while the *xpEHH* indicates
311 selection in the haplometrotic population).

312

313 *Putative function of candidate genes*

314 The F_{st}- and *xpEHH*-based approaches to detect selection yielded a set of 145 candidate genes. To identify
315 putative function of these candidate genes, we performed a BLAST (blastx) (Altschul et al. 1990) search
316 against the nr database. In addition to BLAST, we used OrthoFinder (v. 2.3.1) (Emms and Kelly 2019) to
317 identify *Drosophila melanogaster* orthologs to our candidate genes. To explore whether any of our
318 candidate genes showed differential expression between the two social forms, we performed BLAST
319 searches against *P. californicus* assembled transcripts from Helmkampf et al. (2016) (Supporting Table S5).

320 *Functional annotation of candidate SNPs*

321 We used BEDtools' *intersect* function (v. 2.29.2) (Quinlan and Hall 2010) to extract regions identified
322 simultaneously by the F_{st}- and *xpEHH*-based approach. Next, we extracted all SNPs (a total of 3,782 SNPs)
323 within these regions. We used the program SnpEff (v. 5.0) (Cingolani et al. 2012) to identify the functional
324 role of all SNPs that we recovered from candidate regions identified by our two approaches. We
325 categorized all SNPs based on which genomic feature is affected and whether the SNPs (in coding regions)
326 are synonymous or nonsynonymous substitutions (Table 1).

327 *Haplotype frequencies*

328 We calculated haplotype frequencies for the supergene region and tested for potential departure from
329 Hardy-Weinberg equilibrium using a χ^2 test (Supporting Table S6; Supporting Figure S3). Haplotypes were
330 assigned manually, based on visual inspection of the suggested supergene region (Supporting Table S7);
331 predominantly homozygous for the reference alleles (i.e. from a pleometrotic population) and alternate
332 allele were assigned "Sp/Sp" and "Sh/Sh", respectively, and predominantly heterozygous regions were
333 assigned "Sh/Sp".

334 *RAD sequencing and analysis*

335 Winged males from a monogynous colony in Lake Henshaw (H-population) were collected during the
336 mating season 2019 and immediately stored in 100% ethanol. DNA was isolated using a phenol-chloroform
337 method (Gadau 2009a). Tests using six microsatellite markers confirmed that all males belonged to the
338 same matriline (results not shown). Following Inbar et al. (2020), we constructed a reduced representation
339 genomic library using an adjusted double-digest Restriction-site Associated DNA sequencing (ddRADseq)
340 protocol (Parchman et al. 2012, Peterson et al. 2012). First, we digested DNA withEcoR1 as rare-cutter and
341 Mse1 as frequent-cutter restriction enzyme. Then, we ligated DNA fragments to barcoded adaptors for

342 multiplexing (using CutSmart buffer NEB). The adaptor ligated products were PCR amplified, pooled, and
343 size-selected using Beckman Coulter, Agencourt AMPure XP beads. The libraries were sequenced in one
344 lane of a paired-end, 150 bp reads, on a HiSeq X Ten Illumina sequencer.

345 A total of 369,727,466 paired-reads from the RADseq were examined for quality using fastQC (v. 0.11.7).
346 Trimmomatic v. 0.39 (Bolger et al. 2014) was employed on the raw data to remove adaptor sequences
347 with the following settings: *ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:8:true SLIDINGWINDOW:4:15*
348 *MINLEN:100*. Trimmed reads were then processed by *process_radtags* (v. 2.2) from the Stacks pipeline
349 (Rochette et al. 2019) to demultiplex the data and remove low quality reads as well as those that lacked
350 the enzyme cut sites .

351 Mapping to the genome assembly of *P. californicus* (Pcal2) was performed using BWA-MEM (v. 0.7.17-
352 r1188) with default parameters (Li and Durbin 2009). The resulting alignments were further processed by
353 the *ref_map.pl* and populations available under the STACKS pipeline resulting in 55,340 raw variants. We
354 next used VCFtools (v. 0.1.16) (Danecek et al. 2011) to only keep biallelic markers, with no missing data
355 (i.e. genotyped in all males) and a MAF ≥ 0.45, yielding 4,082 markers. Finally, we excluded heterozygous
356 markers (likely sequencing errors as males are haploid) producing a final set of 2,980 markers used for
357 building genetic maps.

358 The 2,980 markers were all mapped with MultiPoint [http://www.multiqtl.com] using the Backcross
359 population setting. A maximum threshold rfs value of 0.25 resulted in 29 linkage groups. Note, MultiPoint
360 implemented a feature that recognizes markers showing linkage in the opposite phase (to automatically
361 recognize phasing errors in large data sets), hence it is no longer necessary to first double the markers,
362 switch phase for one marker set and then start mapping as described in Gadau (2009b). Multipoint linkage
363 analysis of loci within each linkage group (LG) refined marker order through resampling (jack-knifing,
364 Mester et al. 2003). A skeleton map was generated using delegate markers for all markers that showed
365 100% linkage. Kosambi`s mapping function was used to convert recombination frequencies to
366 centimorgan (Kosambi 1943). The complete map will be published elsewhere, for this manuscript we are
367 only reporting the results for linkage group (LG) 14 that harbors a major non-recombining area
368 (supergene).

369

370 *Characterization of the supergene*

371 To investigate the potential supergene identified in the current study, we used SNP markers identified in
372 the data generated for the queens and the males (RADseq). We used BCFtools' (v. 1.9) *isec* and *merge*
373 functions to extract shared SNPs between the two datasets found on LG14 (Supporting Table S7).

374 The resulting SNP markers (1,064 in total) were used to visualize the genotypes of the males and queens
375 at LG14 using VIVA (v. 0.4.0) (Tollefson et al. 2019). To further characterize the two alleles of the
376 supergene, we performed a PCA on 65 SNP markers identified in the non-recombining area of LG14 (i.e.
377 scaffolds 22, 23, 38, 42 and 48) in datasets of males and queens.

378 To investigate recombination suppression in the supergene region, we used SNP data from six queens
379 homozygous for the Sh allele (H104W, H33G, H34G, H38W, H73W and H98W) and six others homozygous
380 for the Sp allele (P22P, P102R, P114W, P109R, P34R, P84P), assigned based on visual inspection of their
381 genotypes (Supporting Figure S3). We first performed a VCF lift over (on scaffolds 14, 22, 23, 38, 42, 48
382 and 6 into LG14) using flo (Pracana et al. 2017) to generate a chain file and picard's LiftoverVcf tool (v.
383 2.20.0) to perform the lift over. We then used VCFtools (v. 0.1.16) (Danecek et al. 2011) to only keep bi-
384 allelic variants with a MAF ≥ 0.2 and to thin the sites to help plotting and visualization. We next used the
385 R package LDheatmap (v. 1.0-4) (Shin et al. 2006) to generate heatmap plots to visualize pairwise linkage
386 disequilibrium (LD) (Supporting Figure S9).

387

388 **Results**

389 *Pogonomyrmex californicus genome assembly*

390 The new genome assembly (Pcal2) based on pleometrotic *P. californicus* samples spans 252.3 Mb across
391 199 scaffolds (N50 = 10.4 Mb, largest scaffold 20.75 Mb), which is a substantial improvement compared
392 to the previously published assembly for this species (N50 = 0.2 Mb, Bohn et al. 2021). We ran BUSCOv3
393 in genome mode against hymenoptera_odb9 and recovered 98.5% complete BUSCOs
394 (S:97.8%,D:0.7%,F:1.1%,M:0.4%,n:4415), suggesting a nearly complete genome assembly. Genome
395 annotation identified 22.79% repetitive sequence and 15,899 protein coding genes (BUSCO annotation
396 scores C:96.1%[S:90.8%,D:5.3%],F:1.9%,M:2.0%), resembling repeat and gene contents reported for other
397 published ant genomes. More details on the quality of the assembly can be found in the Supplementary
398 Information.

399

400 *The haplometrotic and pleometrotic populations are genetically distinct*

401 To investigate the population structure of the two populations, we conducted a Principle Component
402 Analysis (PCA), based on 314,756 LD-pruned SNPs. PC 1 clearly separates the two populations (Figure 1A),
403 consistent with phylogenetic analyses of the CO1 barcoding data (Supporting Figure S4). Further, the PCA
404 suggests higher genetic diversity in the H-population than the P-population. Interestingly, one queen
405 (P99P) clustered exactly between both populations (Figure 1A), most likely representing an F1 hybrid
406 between a haplo- and a pleometrotic individual from the two different populations, which are only 50 km
407 apart.

408 A detailed analysis of coancestry with fineSTRUCTURE showed that queens share more haplotypes in
409 within-population donor-recipient pairs than in between-population pairs (Figure 1B), consistent with the
410 results obtained by PCA. However, we found that between-population haplotype-sharing was increased
411 in some of the individuals: For example, the putative hybrid queen P99P (and to a lesser extent queen
412 H38W) received a similar number of haplotypes from the haplometrotic and pleometrotic populations.
413 This increased between-population haplotype sharing is consistent with an ongoing admixture between
414 the haplometrotic and pleometrotic populations. In fact, ADMIXTURE analyses (Figure 1B; Supporting
415 Figure S5) supported this view, showing that some queens of the pleometrotic population share ancestry
416 with the other population (e.g. queen P99P, and to a lesser extent queen H33G).

417

418 *Gene flow between the H- and P-populations is a recent phenomenon*

419 To better understand the history and demographic dynamics of the pleometrotic and haplometrotic
420 populations, we used MSMC2 (Schiffels and Wang 2020) (Figure 1C). Both populations showed a similar
421 trajectory with a zigzag-like pattern in effective population size (N_e) between approximately 100,000 to
422 100 generations ago. During this period, both populations experienced two bottlenecks: an ancient one
423 about 100,000 generations ago until about 50,000 generations ago, after which the populations have
424 recovered and a recent one between 4,000 to roughly 500 generations ago. Following the first bottleneck,
425 N_e consistently remained smaller in the pleometrotic population until between 400 and 100 generations
426 ago, when N_e started to increase in both populations, reaching similarly high estimates in recent times.

427 We inferred how populations separated over time by estimating relative cross-coalescence (rCCR,
428 (Schiffels and Durbin 2014)) and found that the two populations started diverging during the first
429 bottleneck at ~100,000 generations ago. Around 1,000 generations ago, both populations were almost
430 completely isolated (rCCR ≈ 0). Between 400 and 100 generations ago, rCCR increased rapidly, suggesting
431 a recent onset of secondary contacts and growing genetic exchange between the two populations,
432 coinciding with population expansions in both populations, as indicated by recent surges in N_e.

433

434 *Genetic diversity is significantly lower in the pleometrotic population*

435 The analyses of population structure also indicated that the P-population is less variable than the H-
436 population. To investigate this further, we estimated nucleotide diversity (π) in 100 kb non-overlapping
437 window across the genome in both populations. Nucleotide diversity was significantly lower in the P-
438 (median = 4.71e-4) compared to the H-population (median = 7.33e-4) (Figure 2A; Wilcoxon test: W =
439 3497100, $p < 2.2e-16$). Similar to π, heterozygosity was significantly lower in the P- (median = 0.167)
440 compared to the H-population (median = 0.274) (Figure 2B; Wilcoxon test: W = 288, p = 4.507e-07).
441 Consistent with this observation, inbreeding was highly prevalent in the P-population (median = 0.341) but
442 absent in the H-population (median = -0.083) (Figure 2C; Wilcoxon test: W = 16, p = 4. 507e-07). Such
443 reductions of diversity in the pleometrotic population could be explained by a recent bottleneck (founder
444 effect) followed by inbreeding and/or purifying selection.

445 To investigate deviations from neutrality, we estimated Tajima's *D* in 100 kb non-overlapping windows in
446 the two studied populations (Figure 2D). Tajima's *D* is a powerful test based on the difference between π
447 and its expected value, commonly employed in population genetics to detect selection, but also to detect
448 recent demographic processes (Tajima 1989). Tajima's *D* was positive and differed slightly between the
449 two populations (median $_{haplometrotic}$ = 1.09; median $_{pleometrotic}$ = 1.08; W = 2110726, p = 0.005). Interestingly,
450 average Tajima's *D* was significantly greater than 0 in the pleometrotic (one-sided t-test; mean = 1.01; t =
451 71.862, df = 2003, p < 2.2e-16) and haplometrotic (one-sided t-test; mean = 1.11; t = 177.68, df = 2003, p
452 < 2.2e-16) populations suggesting ongoing population contraction.

453 Since inbreeding is prevalent in the pleometrotic population, we expected to find a lower linkage
454 disequilibrium (LD) decay in this population. However, mean LD estimates dropped rapidly to half its
455 original value in both populations along the first ~700 bp (P-population) and ~500 bp (H-population),
456 reflecting the high recombination rates in social insects (Figure 2E). However, mean LD estimates in the
457 pleometrotic populations were overall higher (r^2 = 0.7 [0.1 kb] to r^2 = 0.19 [5 kb]) than in the haplometrotic
458 population (r^2 = 0.57 [0.1 kb] to r^2 = 0.17 [5 kb]), suggesting a reduced recombination rate and/or genetic
459 variation in the pleometrotic population.

460 *The social polymorphism in P. californicus is a polygenic trait involving a supergene*

461 Our genome scans uncovered large genomic spans on two scaffolds (22 and 23) showing significant
462 deviations from the rest of the genome throughout their entire lengths. Both scaffolds have high levels of
463 differentiation, reduced nucleotide diversity, a negative Tajima's *D*, as well as very high *xpEHH* scores in
464 the pleometrotic population (Supporting Figure S6). Further analyzes revealed that SNP markers are in
465 high LD, with little to no LD decay in these scaffolds (Supporting Figure S7). Similarly, visual inspection of
466 the genotypes of all 35 queens showed little or no recombination in these scaffolds (Supporting Figure S3),
467 similar to the patterns described for supergenes in the ants *Solenopsis* and *Formica*.

468 To investigate this further, we conducted a linkage analysis on 108 haploid males from a monogynous
469 colony from the H-population using 2,980 RAD markers. Note, in Hymenoptera males develop from
470 unfertilized eggs due to the haplodiploid sex determination. There were no missing data and no marker
471 showed a significant segregation distortion. Multipoint mapping put all markers in 29 Linkage groups (data
472 not shown).

473 Linkage group 14, containing the putative supergene region, covers 255 cM (~26 Mb) and 339 RAD markers
474 (Supporting Figure S8). The non-recombining region is embedded solidly in the middle of a chromosome,
475 spanning across five complete physical scaffolds (22, 23, 38, 42, 48), and parts of scaffolds 6 and 14
476 (Supporting Figure S8; Supporting Table S7). Visual inspection of markers from both males and queens
477 showed that haplotypes extended along the entire supergene region, indicating absence of recombination
478 (Figure 3A). The putative supergene covered a minimum of ~8 Mb, based on the non-recombining region
479 in linkage group 14.

480 PCA on 65 markers along the non-recombining region showed that PC1, which explains 97 % of the
481 variation, separates the male dataset into two distinct clusters (Sh and Sp alleles), indicating that these
482 males are the progeny of a heterozygous Sh/Sp mother queen (Figure 3B). Adding queens from the P- and
483 H-population to the PCA revealed that all but two pleometrotic queens were homozygous for the Sp allele,
484 while most haplometrotic queens were either heterozygous (Sh/Sp) or homozygous for the Sh allele
485 (Figure 3B).

486 Recombination suppression between Sh and Sp was further confirmed by examining LD across LG14 using
487 the genomes of Sh and Sp queens (Supporting Figure S9). LD along the putative supergene is reduced in
488 the Sh/Sp heterozygous (Supporting Figure S9A), but not among homozygous Sh/Sh (Supporting Figure
489 S9B) and Sp/Sp queens (Supporting Figure S9C), indicating recombination in homozygous, but not in
490 heterozygous queens. Such absence of recombination predominantly between heterozygotes is a common
491 characteristic of supergenes (Wang et al. 2013; Schwander et al. 2014; Charlesworth 2016; Yan et al. 2020).

492 Unlike other social supergene systems (e.g. in *Solenopsis*) (Ross et al. 2007), there were no deviations from
493 Hardy-Weinberg equilibrium regarding the frequency of Sh/Sh, Sh/Sp and Sp/Sp genotypes at the
494 supergene locus in either population (pleometrotic: $\chi2$ (df = 1, N = 19) = 0.1396, p = 0.713; haplometrotic:
495 $\chi2$ (df = 1, N = 16) = 0. 0.0711, p = 0.789) (Supplemental Table S6).

496

497 Genetic differentiation between the H- and P-population is significantly increased in the non-recombining
498 region compared to the remainder of LG14 (Figure 3C; Supporting Figure S10; Wilcoxon test: W = 12261,
499 p < 2.2e-16). Nucleotide diversity and Tajima's *D* on the other hand are significantly reduced in the
500 pleometrotic population (Figure 3C; Supporting Figure S10; Kruskal–Wallis rank sum test $_{\text{nucleotide diversity}}$, $\chi2$

501 = 254.77, df = 3, p < 2.2e-16; Kruskal–Wallis rank sum test $_{\text{Tajima's } D}$, $\chi 2$ = 177.81, df = 3, p < 2.2e-16; results
502 of pairwise Wilcoxon rank sum *post hoc* tests are displayed in the figure).

503 The supergene region includes 682 (4.3 %) of the 15,899 annotated *P. californicus* genes. Ten out of the
504 682 genes were previously shown to be differentially expressed between aggressive and non-aggressive
505 queens, while another ten genes showed differential expression depending on the social context of the
506 queens (Supporting Table S8) (Helmkampf et al. 2016).

507 As a typical feature of supergenes, transposable element (TE) content in the non-recombining region was
508 significantly increased compared to the rest of LG14 (Supporting Figure S11B; Kruskal–Wallis rank sum
509 test, $\chi 2$ = 342.74, df = 3, p < 2.2e-16; results of pairwise Wilcoxon rank sum *post hoc* tests are displayed in
510 the figure). Exon content on the other hand was reduced in this region (Supporting Figure S11A).

511 In addition to the supergene region, our genome scans revealed a number of regions showing signatures
512 of selection in the pleometrotic population. These highly differentiated regions that are likely under
513 divergent selection were identified by estimating F_{st} in 100 kb non-overlapping windows across the
514 genome. Windows were only considered as significantly differentiated if they were in the top 5% tail of
515 the distribution (Figure 4A). We used estimates of Tajima's *D* and nucleotide diversity in 100 kb non-
516 overlapping windows to determine in which population selection acted on the differentiated regions
517 (Supporting Figure S6). After clustering adjacent windows, we retained 51 candidate regions, of which 49
518 showed signatures of selection in the P-population (Supporting Table S2).

519 Further, we performed a cross-population extended haplotype homozygosity (*xpEHH*) (Sabeti et al. 2007)
520 test to screen the populations for selective sweeps using phased data of unrelated queens (Supporting
521 Figure S6). The *xpEHH* analysis yielded a total of 340 outlier SNPs and a total of 32 genomic regions under
522 selection (Supporting Tables S3).

523 Overlapping both analyses resulted in 17 remaining candidate regions under selection in the P-population
524 (Supporting Tables S4). Exonic SNPs in all 17 candidate regions included 99 non-synonymous variants
525 (Table 1), suggesting selection at the protein level in some of our candidate genes. We also find 84
526 synonymous changes that are "silent" at the protein level, but that can also be targeted by selection
527 (Cuevas et al. 2012).

528 Intriguingly, the supergene contains 12 of the 17 regions (on scaffolds 14, 22 and 23) that we identified to
529 evolve under divergent selection between the two populations.

530 The five remaining regions showing clear signatures of selective sweeps in the P-population are located
531 on scaffolds 1, 9, 15, and 16 (Figure 4B).

532 Among the 57 genes in these five regions, we find two particularly interesting candidates showing
533 signatures of a hard selective sweep in the pleometrotic population (i.e. high F_{st}, negative Tajima's *D*, and
534 a peak of *xpEHH*), a *histone acetyltransferase MYST2* and a *DNMT1* gene found in differentiated regions
535 on scaffold 15 and 16, respectively (Figure 4B; Supporting Table S5). The two SNPs associated with the
536 highest *xpEHH* peaks found in these regions on scaffold 15 and 16 show a clear signature of site-specific
537 extended haplotype homozygosity (*EHHS*) in the pleometrotic, but not the haplometrotic population
538 (Figure 4C). Visual inspection of variant positions in *MYST2* and *DNMT1* showed that all pleometrotic
539 queens are fixed for the reference allele (with the exception of the hybrid P99P queen), while
540 haplometrotic queens are either heterozygous or homozygous for alternate alleles (Supporting Figures
541 S12-S13), consistent with a selective sweep fixing the reference allele in the pleometrotic population.
542 Other notable genes include an ortholog of *D. melanogaster Gem2* gene involved in motor behavior
543 (Cauchi et al. 2010; Borg and Cauchi 2013) as well as an ortholog of *Neurexin 1*, involved in the regulation
544 of learning and memory, locomotion and visual function (Li et al. 2007; Zeng et al. 2007) (Supporting Table
545 S5).

546

547 **Discussion**

548 The social niche polymorphism in colony founding of *P. californicus* was first reported almost 20 years ago
549 (Rissing et al. 2000). It eventually leads to fundamentally different colony structures, where either
550 unrelated queens cooperate (i.e. primary polygyny) or a single queen monopolizes all reproduction (i.e.
551 primary monogyny, Overson et al. 2016). Since its original description, we have learned a lot about
552 behavioral, ecological, and population genetic correlates of these complex alternative phenotypes
553 (Johnson 2002; Johnson 2004; Clark and Fewell 2014; Overson et al. 2014; Overson et al. 2016; Shaffer et
554 al. 2016; Haney and Fewell 2018). From these studies it has emerged that the core difference generating
555 these two different social niches is the evolution of tolerance for additional co-founding queens. However,
556 the underlying molecular mechanisms (but see Helmkampf et al. 2016), the genetic architecture, and
557 evolutionary dynamics remain largely unknown.

558 Our study reveals that the haplometrosis/pleometrosis polymorphism in *P. californicus* is associated with
559 an approximately 8 Mb non-recombining region, reminiscent of supergene loci described for
560 polygyny/monogyny polymorphisms in *Formica* wood ants and *Solenopsis* fire ants (Wang et al. 2013;
561 Purcell et al. 2014). However, we also find genomic, population genetic, and ecological characteristics
562 unique to the *Pogonomyrmex* system that highlight the independent evolutionary histories of the social
563 niche polymorphism in *Pogonomyrmex*, *Formica* and *Solenopsis*.

564 Our analyses provide several lines of evidence for the presence of a supergene on LG14 associated with
565 the expression of haplometrotic or pleometrotic behavior. According to estimates of F_{st}, Tajima's *D*, and
566 haplotype homozygosity, genetic differentiation between individuals from the two social forms is
567 significantly higher in this region compared to most of the remaining genome (Figure 3C; Supporting Figure
568 S6). Pleometrotic queens are predominantly homozygous (84.2 %) for a single haplotype (Sp) in this region,
569 while haplometrotic queens carried either two alternative haplotypes Sh/Sp (50%), were homozygous for
570 the alternative allele (Sh/Sh) (37.5 %), or were homozygous for Sp/Sp (12.5%). Very low LD decay and
571 absence of recombinant F1 offspring at this locus suggest little to no recombination between Sh and Sp
572 haplotypes (Supporting Figure S9). This region contains 88 annotated genes that show both extended
573 homozygosity and outlier F_{st} (Supporting Table S5) and is significantly enriched in TEs (Supporting Figure
574 S11).

575 Together, these data clearly suggest the presence of an ~8 Mb supergenic region of linked genetic loci on
576 one chromosome of *P. californicus* that is associated with the social polymorphism in this species, i.e. a
577 social supergene.

578 Apart from the social supergene, we identify five other differentiated genomic regions (physically
579 unlinked) between haplo- and pleometrotic populations. This suggests that the social niche polymorphism
580 could in addition to the supergene be influenced by other epistatic or additive interactions with these
581 regions. The regions contain 57 genes that show both extended homozygosity and outlier F_{st} (Supporting
582 Table S5, range 8-22 genes), including the DNA methyltransferase *DNMT1* and a histone acetyltransferase
583 (*MYST2*) gene that could be involved in epigenetically regulating the expression of the alternative
584 phenotypes. The presence of unlinked modifier loci influencing the expression of major quantitative traits
585 is not unusual (Flint and Mackay 2009), yet none have been described so far for the known social
586 supergenes.

587 Several other characteristics of the *P. californicus* social polymorphism differ from the well-studied
588 examples in *Solenopsis* and *Formica.* In both *Solenopsis* and *Formica*, there is little evidence for genetic
589 differentiation of polygyne and monogyne populations, with both forms often occurring in sympatry
590 (Chapuisat et al. 2004; Shoemaker et al. 2006). In *P. californicus* however, certain populations are nearly
591 fixed for either the haplo- or the pleometrotic behavior (Overson et al. 2016). The current study further
592 corroborated Overson et al (2016) significantly reduced gene flow between the two focal populations both
593 at the nuclear and mitochondrial level (Figures 1A and 1B; mean genome wide F_{st} = 0.236 (Figure 4A);
594 mitochondrial substructure (Supporting Figure S4)). Genetic differentiation is hence significantly higher
595 than between the two different social forms in *S. invicta* and *Formica selysi*, both within sympatric and
596 between allopatric populations (Chapuisat et al. 2004; Shoemaker et al. 2006). Similar levels of
597 differentiation have however been described for two social forms in *F. exsecta* (Brelsford et al. 2020).

598 Intriguingly, analysis of Sh and Sp haplotype frequencies in *P. californicus* showed no significant deviation
599 from Hardy-Weinberg equilibria in either population, consistent with a lack of intrinsically lethal haplotype
600 combinations. In both other supergenes associated with social polymorphisms in *Formica* and fire ants,
601 such lethal combinations exist and are presumed crucial for the maintenance of the social niche
602 polymorphism in these species (Ross 1997; Avril et al. 2020).

603 In addition, and unlike in *Solenopsis*, the *Pogonomyrmex* supergene genotype is not perfectly correlated
604 with the observed queen phenotype/colony founding type, in that three pleometrotic individuals were

605 heterozygous (Sh/Sp) and two haplometrotic ones were homozygous (Sp/Sp). Nevertheless, colony
606 founding behavior of all 35 sequenced queens produced only one queen that showed a different
607 phenotype than expected in our standardized sandwich assays given the population they came from, i.e.
608 an haplometrotic queen paired with another haplometrotic queen was still alive at the end of our
609 behavioral experiments (Clark and Fewell 2014; Overson et al. 2014; column phenotype, Supporting Table
610 S1). However, while aggression between haplometrotic queens is generally high, not all queens from the
611 haplometrotic population kill each other within 7 days and *vice versa* not all queens from the pleometrotic
612 population are tolerant against co-foundresses (Clark and Fewell 2014; Overson et al. 2014). These
613 observations are consistent with a polygenic architecture of the pleometrosis/haplometrosis social morph
614 polymorphism, with the existence of several co-evolved, but unlinked genetic loci that contribute to the
615 expression of different social phenotypes. Intriguingly, correlation of genotype and phenotype were in fact
616 perfect for the two candidate genes *MYST2* and *DNMT1* that lie outside of the supergene (Supporting
617 Figures S12 and S13). Environmental conditions, in particular the social context and interactions with co-
618 founding queens (indirect genetic effects), might further shift the expression of certain behavioral traits
619 involved in this complex social niche polymorphism (Clark and Fewell 2014; Overson et al. 2014, Saltz et
620 al. 2016) .

621 The core behavioral difference between aggressive/haplometrotic queens and tolerant/pleometrotic
622 queens is not that pleometrotic queens are intrinsically docile but that their propensity for reciprocal
623 aggression towards aggressive co-foundresses is low and their ability to adjust their behavioral response
624 according to the social environment is high (Clark and Fewell 2014; Overson et al. 2014) . That is in stark
625 contrast to haplometrotic queens that appear essentially fixed in their behavioral response towards other
626 queens. Similarly, haplometrotic queens show less behavioral plasticity also in other traits such as egg
627 laying or excavation (Clark and Fewell 2014).

628 Our coalescent analysis indicates that the haplo- and pleometrotic populations started to separate
629 approximately 100,000 generations ago, reaching almost complete isolation around 1,000 generations
630 ago, before gene flow between those two populations increased again. In accordance, some individuals
631 showed clear patterns of hybridization, recent co-ancestry and admixture (e.g. potential F1 hybrid queen
632 P99P; Figures 1A and 1B), consistent with migration between the two populations, which are only 50 km
633 apart. Migration might predominantly occur from the haplometrotic into the pleometrotic population, as
634 indicated by the analysis of haplotype sharing across populations (Figure 1B) and the occurrence of a

635 potentially hybrid individual in the pleometrotic population. The onset of gene flow between the
636 haplometrotic and pleometrotic population starting ~400 generations ago could secondarily reduce
637 genetic differentiation between both forms everywhere but in the non-recombining supergene, eventually
638 resulting in the pattern also observed in *Solenopsis invicta* and *Formica selysi*, where genetic
639 differentiation between both social morphs nearly exclusively occurs in the supergene (Purcell et al. 2014;
640 Pracana et al. 2017).

641 The demographic and population genetic analyses are consistent with our current understanding of
642 pleometrosis to have evolved as a local adaptation to ephemeral, small and overall harsher habitats in the
643 mountains (Sierra Nevada, California, USA), where cooperative colony founding is beneficial (Haney and
644 Fewell 2018) and can reach high frequencies. The pleometrotic population showed a reduced genetic
645 variability and higher inbreeding coefficient (median F_{is} = 0.34, SD = 0.23 for pleometrotic, median F_{is} = -
646 0.08, SD = 0.06 for haplometrotic population) compared to the haplometrotic population (Figures 2A-C).
647 A recent bottleneck during colonization could explain this overall reduced genetic diversity and higer
648 inbreeding in this population, whereas the occurrence of admixture argues against isolation as a major
649 factor of inbreeding and reduced genetic diversity. Haplometrosis is considered the dominant social form
650 in most populations of the species (Johnson 2002; Johnson 2004; Overson et al. 2014), which is in
651 alignment with higher nucleotide diversity and heterozygosity in our focal haplometrotic population.

652 Supergenes have long been suspected to facilitate the emergence and maintenance of complex
653 phenotypes (Schwander et al. 2014). The success of the polygynous form in *S. invicta* in its introduced
654 range is attributed to its ability to efficiently colonize disturbed areas (King and Tschinkel 2006) rather than
655 its higher competitive abilities. This might by similar in *P. californicus* where pleometrosis can be very
656 successful (reaching fixation) in newly colonized recently disturbed habitats. The observation that both
657 populations show genome-wide positive Tajima's *D* (Figure 2D) suggests a population contraction. The
658 habitat preference of *P. californicus* for disturbed areas and the observation that populations vanish if
659 disturbed areas are overgrown, supports the idea that both social types encounter regularly local
660 extinction and recolonization events which could explain the genome wide positive Tajima's *D* values.

661 Theoretical models suggest that pleometrotic populations might be stable against invasions from
662 haplometrotic individuals, if pleometrosis occurs at a certain frequency, because single haplometrotic
663 queens would be out-competed by polygynous colonies (Shaffer et al. 2016). Conversely, the founding
664 success of pleometrotic queens is likely negatively correlated with the frequency of haplometrotic

665 individuals in a population, as pleometrotic/haplometrotic pairings commonly result in the death of the
666 former. This would explain the rareness of high frequency pleometrotic populations (Johnson 2002), as
667 such populations can only evolve when the frequency of pleometrosis constantly remains above a certain
668 threshold.

669 Apart from the supergene, we found additional genomic regions showing signatures of a selective sweep
670 in the pleometrotic population. Functional characterization of the genes contained in these regions
671 uncovered several candidate genes that might be particularly relevant for the expression of haplo- and
672 pleometrosis. Among others, these candidate regions include *DNMT1* and an *histone acetyltransferase*
673 (showing non-synonymous substitutions, Supporting Figures S12-S13) required for epigenetic
674 modifications via DNA methylation or histone modification. Further studies are necessary to decipher,
675 whether alternate alleles of these genes indeed contribute to the expression of haplometrotic or
676 pleometrotic behavior during colony founding in *P. californicus*.

677

678 **Conclusion**

679 Pleometrosis and primary polygyny in *P. californicus* appear to have evolved as an adaptation to harsh
680 environments, allowing higher colony/queen densities and greater colony productivity under resource
681 limitation. However, this social niche is generally rare and has reached fixation in only a few populations.
682 We describe a new supergene contributing this social niche polymorphism, showing similarities but also
683 remarkable differences to previously studied social supergenes in *Formica* and *Solenopsis*. The
684 *Pogonomyrmex* supergene is similar in size (8 Mb vs. 11 and 12.7 Mb) and we identified two major alleles
685 segregating in the population. So far, there is no evidence for intrinsic lethality or the selective killing of
686 individuals carrying a certain genotype combination, but we cannot exclude such mechanisms to exist at
687 this point. Whereas the polygynous queens are heterozygous for the supergene in *Solenopsis* (SB/Sb) and
688 *Formica* (Sm/Sp), polygynous/pleometrotic queens in *P. californicus* are predominantly homozygous
689 (Sp/Sp). Our initial analyses suggest that the haplometrotic Sh allele is dominant, as both Sh/Sh and Sh/Sp
690 haplometrotic queens occur in the H-population. This might limit the frequency of pleometrosis, but it also
691 allows for its persistence in an otherwise predominantly haplometrotic species as a recessive allele
692 invisible to selection. Apart from the supergene, there are at least five additional regions showing
693 signatures of a selective sweep associated with the social polymorphism. Further studies are necessary to
694 understand the relevance of these loci for the expression of the different social phenotypes as they might
695 turn out to be more important than the supergene itself. Mixed populations of haplo- and pleometrotic *P.*
696 *californicus* are well known, offering exciting avenues to further explore evolutionary dynamics of social
697 niche polymorphism in this species. Our study showcases that the genetic architecture underlying social
698 polymorphism might be much more complex and diverse than previously appreciated.

699

700 **Acknowledgements**

701 We thank Pine Valley Academy (Pine Valley, California, USA) for support of the fieldwork and the staff of
702 Lake Henshaw resort (Lake Henshaw, California, USA) for permission to collect ants on their premises.
703 Many thanks to Jenny Märzhäuser, Tobias van Elst and Ti Eriksson for help in collecting the ants, and to
704 Jennifer Fewell for hosting us at Arizona State University, Tempe, Arizona, USA. We are grateful for Hilde
705 Schwitte's excellent help in DNA extraction, barcoding, and library construction, and to Una Hadziomerovic
706 for extracting DNA for the genome assembly. We acknowledge the service of the Core Facility Genomik
707 (CFG) at Münster University, Germany. This research and UE was funded by the German Research
708 Foundation (DFG) as part of the SFB TRR 212 (NC^3) – TP C04 project numbers 316099922 and 396780988.
709 ME was funded by the German Research Foundation (DFG) – 403813881 with grant to L.S. (SCHR 1554/2-
710 1) under the priority program "Rapid evolutionary adaptation: Potential and constraints" (SPP 1819). ME
711 acknowledges support by the DFG Research Training Group 2220 "Evolutionary Processes in Adaptation
712 and Disease".

References

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, Madeoy J, Nicholas TJ, Neff MW. 2010. Tracking footprints of artificial selection in the dog genome. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 107:1160–1165. Available from: www.pnas.org/cgi/doi/10.1073/pnas.0909918107

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* [Internet] 19:1655–1664. Available from: /pmc/articles/PMC2752134/?report=abstract

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* [Internet] 43:11.10.1-11.10.33. Available from: https://currentprotocols.onlinelibrary.wiley.com/doi/full/10.1002/0471250953.bi1110s43

Avril A, Purcell J, Béniguel S, Chapuisat M. 2020. Maternal effect killing by a supergene controlling ant social organization. *Proc. Natl. Acad. Sci. U. S. A.* 117:17130–17134.

Bernasconi G, Strassmann JE. 1999. Cooperation among unrelated individuals: The ant foundress case. *Trends Ecol. Evol.* 14:477–482.

Bohn J, Halabian R, Schrader L, Shabardina V, Steffen R, Suzuki Y, Ernst UR, Gadau J, Makałowski W. 2021. Genome assembly and annotation of the California harvester ant *Pogonomyrmex californicus*.Betran E, editor. *G3 Genes|Genomes|Genetics* [Internet] 11. Available from: https://academic.oup.com/g3journal/article/doi/10.1093/g3journal/jkaa019/6020260

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* [Internet] 30:2114–2120. Available from: https://pubmed.ncbi.nlm.nih.gov/24695404/

Boomsma JJ, Huszár DB, Pedersen JS. 2014. The evolution of multiqueen breeding in eusocial lineages with permanent physically differentiated castes. *Anim. Behav.* 92:241–252.

Borg R, Cauchi RJ. 2013. The Gemin Associates of Survival Motor Neuron Are Required for Motor Function in Drosophila.Gilestro GF, editor. *PLoS One* [Internet] 8:e83878. Available from: https://dx.plos.org/10.1371/journal.pone.0083878

Brelsford A, Purcell J, Avril A, Tran Van P, Zhang J, Brütsch T, Sundström L, Helanterä H, Chapuisat M. 2020. An Ancient and Eroded Social Supergene Is Widespread across Formica Ants. *Curr. Biol.* [Internet] 30:304-311.e4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0960982219315027

Buckley SB. 1867. Descriptions of new species of North American Formicidae. Available from: https://zenodo.org/record/27047

Cauchi RJ, Sanchez-Pulido L, Liu JL. 2010. Drosophila SMN complex proteins Gemin2, Gemin3, and Gemin5 are components of U bodies. *Exp. Cell Res.* 316:2354–2364.

Chang C, Chow C, Tellier LC, Vattikuti S, Purcell S, Lee J. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* [Internet] 4:7. Available from: https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533

Chapuisat M, Bocherens S, Rosset H. 2004. Variable queen number in ant colonies: No impact on queen turnover, inbreeding, and population genetic differentiation in the ant Formica selysi. *Evolution (N. Y).* 58.

Charlesworth D. 2016. The status of supergenes in the 21st century: recombination suppression in

758 <scp>B</scp> atesian mimicry and sex chromosomes and other complex adaptations. *Evol. Appl.*
759 [Internet] 9:74–90. Available from: https://onlinelibrary.wiley.com/doi/10.1111/eva.12291
760 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for
761 annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the
762 genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* [Internet] 6:80–92.
763 Available from: https://pubmed.ncbi.nlm.nih.gov/22728672/
764 Clark RM, Fewell JH. 2014. Social dynamics drive selection in cooperative associations of ant queens.
765 *Behav. Ecol.* [Internet] 25:117–123. Available from:
766 https://academic.oup.com/beheco/article/25/1/117/223077
767 Cuevas JM, Domingo-Calap P, Sanjuán R. 2012. The fitness effects of synonymous mutations in DNA and
768 RNA viruses. *Mol. Biol. Evol.* [Internet] 29:17–20. Available from:
769 https://pubmed.ncbi.nlm.nih.gov/21771719/
770 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT,
771 Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* [Internet] 27:2156–
772 2158. Available from: http://samtools.sourceforge.net
773 Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands of genomes.
774 *Nat. Methods* [Internet] 9:179–181. Available from: http://www.wtccc.org.uk/.
775 Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas
776 MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-
777 generation DNA sequencing data. *Nat. Genet.* 43:491–501.
778 Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics.
779 *Genome Biol.* [Internet] 20:238. Available from:
780 https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y
781 Fabian DK, Kapun M, Nolte V, Kofler R, Schmidt PS, Schlötterer C, Flatt T. 2012. Genome-wide patterns of
782 latitudinal differentiation among populations of Drosophila melanogaster from North America. *Mol.*
783 *Ecol.* 21:4748–4769.
784 Flint J, Mackay TFC. 2009. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome*
785 *Res.* 19.
786 Gadau J. 2009a. DNA isolation from ants. *Cold Spring Harb. Protoc.* [Internet] 4:pdb.prot5245. Available
787 from: http://cshprotocols.cshlp.org/content/2009/7/pdb.prot5245.full
788 Gadau J. 2009b. Phase-unknown linkage mapping in ants. *Cold Spring Harb. Protoc.* 4.
789 Gautier M, Klassmann A, Vitalis R. 2017. rehh 2.0: a reimplementation of the R package rehh to detect
790 positive selection from haplotype structure. In: Molecular Ecology Resources. Vol. 17. Blackwell
791 Publishing Ltd. p. 78–90. Available from: http://doi.wiley.com/10.1111/1755-0998.12634
792 Glancey BM, Lofgren CS. 1988. Adoption of Newly-Mated Queens: A Mechanism for Proliferation and
793 Perpetuation of Polygynous Red Imported Fire Ants, Solenopsis invicta Buren. *Florida Entomol.*
794 71:581.
795 Haney BR, Fewell JH. 2018. Ecological drivers and reproductive consequences of non-kin cooperation by
796 ant queens. *Oecologia* [Internet] 187:643–655. Available from: https://doi.org/10.1007/s00442-
797 018-4148-9
798 Helmkampf M, Mikheyev AS, Kang Y, Fewell J, Gadau J. 2016. Gene expression and variation in social
799 aggression by queens of the harvester ant Pogonomyrmex californicus. *Mol. Ecol.* [Internet]
800 25:3716–3730. Available from: http://doi.wiley.com/10.1111/mec.13700
801 Helms KR, Newman NJ, Helms Cahan S. 2013. Regional variation in queen and worker aggression in
802 incipient colonies of the desert ant Messor pergandei. *Behav. Ecol. Sociobiol.* [Internet] 67:1563–
803 1573. Available from: https://link.springer.com/article/10.1007/s00265-013-1568-8

Herbers JM. 1986. Nest site limitation and facultative polygyny in the ant Leptothorax longispinosus. *Behav. Ecol. Sociobiol.* [Internet] 19:115–122. Available from: https://link.springer.com/article/10.1007/BF00299946

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags.Begun DJ, editor. *PLoS Genet.* [Internet] 6:e1000862. Available from: https://dx.plos.org/10.1371/journal.pgen.1000862

Hölldobler B, Wilson EO. 1977. The number of queens: An important trait in ant evolution. *Naturwissenschaften* 64.

Inbar S, Cohen P, Yahav T, Privman E. 2020. Comparative study of population genomic approaches for mapping colony-level traits.Mikheyev A, editor. *PLOS Comput. Biol.* [Internet] 16:e1007653. Available from: https://dx.plos.org/10.1371/journal.pcbi.1007653

Johnson RA. 2002. Semi-claustral colony founding in the seed-harvester ant Pogonomyrmex californicus: A comparative analysis of colony founding strategies. *Oecologia* [Internet] 132:60–67. Available from: http://rsb.

Johnson RA. 2004. Colony founding by pleometrosis in the semiclaustral seed-harvester ant Pogonomyrmex californicus (Hymenoptera: Formicidae). *Anim. Behav.* 68:1189–1200.

Katoh K, Rozewicki J, Yamada KD. 2018. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* [Internet] 20:1160–1166. Available from: https://pubmed.ncbi.nlm.nih.gov/28968734/

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* [Internet] 16:980–989. Available from: www.genome.org/cgi/doi/10.1101/gr.5157306.

King JR, Tschinkel WR. 2006. Experimental evidence that the introduced fire ant, Solenopsis invicta, does not competitively suppress co-occurring ants in a disturbed habitat. *J. Anim. Ecol.* 75.

Kolaczkowski B, Kern AD, Holloway AK, Begun DJ. 2011. Genomic differentiation between temperate and tropical Australian populations of Drosophila melanogaster. *Genetics* [Internet] 187:245–260. Available from: www.genetics.org/

Kosambi DD. 1943. The Estimation of Map Distances from Recombination Values. *Ann. Eugen.* 12:172–175.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of Population Structure using Dense Haplotype Data. *PLoS Genet* [Internet] 8:1002453. Available from: http://www.paintmychromosomes.com/.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* [Internet] 25:1754–1760. Available from: http://maq.sourceforge.net

Li J, Ashley J, Budnik V, Bhat MA. 2007. Crucial Role of Drosophila Neurexin in Proper Active Zone Apposition to Postsynaptic Densities, Synaptic Growth, and Synaptic Transmission. *Neuron* [Internet] 55:741–755. Available from: http://www.cell.com/article/S0896627307006137/fulltext

Matschiner M, Maria J, Barth I, Tørresen OK, Star B, Tessand Baalsrud H, Servane M, Brieuc O, Pampoulie C, Bradbury I, et al. 2021. Origin and fate of supergenes in Atlantic cod. *bioRxiv* [Internet]:2021.02.28.433253. Available from: https://doi.org/10.1101/2021.02.28.433253

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* [Internet] 20:1297–1303. Available from: http://www.genome.org/cgi/doi/10.1101/gr.107524.110.

Mester D, Ronin Y, Minkov D, Nevo E, Korol A. 2003. Constructing Large-Scale Genetic Maps Using an Evolutionary Strategy Algorithm. *Genetics* 165.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, Ffrench-Constant RH, Blaxter ML, et al. 2012. Genomic islands of divergence in hybridizing heliconius butterflies identified by large-scale targeted sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* [Internet] 367:343–353. Available from: http://tiling.gersteinlab.

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* [Internet] 32:292–294. Available from: https://academic.oup.com/bioinformatics/article/32/2/292/1744356

Overson R, Fewell JH, Gadau J. 2016. Distribution and origin of intraspecific social variation in the California harvester ant Pogonomyrmex californicus. *Insectes Soc.* [Internet] 63:531–541. Available from: https://asu.pure.elsevier.com/en/publications/distribution-and-origin-of-intraspecific-social-variation-in-the-

Overson R, Gadau J, Clark RM, Pratt SC, Fewell JH. 2014. Behavioral transitions with the evolution of cooperative nest founding by harvester ant queens. *Behav. Ecol. Sociobiol.* [Internet] 68:21–30. Available from: https://link.springer.com/article/10.1007/s00265-013-1618-2

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. *Mol. Ecol.* 21:2991–3005.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7.

Pracana R, Priyam A, Levantis I, Nichols RA, Wurm Y. 2017. The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. *Mol. Ecol.* [Internet] 26:2864–2879. Available from: http://doi.wiley.com/10.1111/mec.14054

Purcell J, Brelsford A, Wurm Y, Perrin N, Chapuisat M. 2014. Convergent genetic architecture underlies social organization in ants. *Curr. Biol.* 24:2728–2732.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* [Internet] 81:559–575. Available from: /pmc/articles/PMC1950838/?report=abstract

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* [Internet] 26:841–842. Available from: /pmc/articles/PMC2832824/?report=abstract

R Core Team. 2020. A Language and Environment for Statistical Computing. 4.0.2. Vienna Available from: http://www.r-project.org/

Rissing SW, Johnson RA, Martin JW. 2000. Colony Founding Behavior of Some Desert Ants: Geographic Variation in Metrosis. *Psyche A J. Entomol.* 103:95–101.

Rochette NC, Rivera-Colón AG, Catchen JM. 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol. Ecol.* 28.

Rosengren R, Sundström L, Fortelius W. 1993. Monogyny and polygyny of Formica ants: a result of alternative dispersal tactics? :308–333. Available from: https://researchportal.helsinki.fi/en/publications/monogyny-and-polygyny-of-formica-ants-a-result-of-alternative-dis

Ross KG. 1997. Multilocus evolution in fire ants: Effects of selection, gene flow and recombination. *Genetics* 145.

Ross KG, Krieger MJB, Keller L, Shoemaker DD. 2007. Genetic variation and structure in native populations of the fire ant Solenopsis invicta: evolutionary and demographic implications.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* [Internet] 449:913–918. Available from:

https://www.nature.com/articles/nature06250
Saltz JB, Geiger AP, Anderson R, Johnson B, Marren R. 2016. What, if anything, is a social niche? *Evol. Ecol.* [Internet] 30:349–364. Available from: https://link.springer.com/article/10.1007/s10682-015-9792-5
Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet. Vol.* 46.
Schiffels S, Wang K. 2020. MSMC and MSMC2: The multiple sequentially Markovian coalescent. In: Methods in Molecular Biology. Vol. 2090. Humana Press Inc. p. 147–166. Available from: https://doi.org/10.1007/978-1-0716-0199-0_7,
Schwander T, Libbrecht R, Keller L. 2014. Supergenes and complex phenotypes. *Curr. Biol.* 24:R288–R294.
Seppä P, Gyllenstrand N, Corander J, Pamilo P. 2004. COEXISTENCE OF THE SOCIAL TYPES: GENETIC POPULATION STRUCTURE IN THE: ANT FORMICA EXSECTA. *Evolution (N. Y).* [Internet] 58:2462–2471. Available from: http://doi.wiley.com/10.1111/j.0014-3820.2004.tb00875.x
Shaffer Z, Sasaki T, Haney B, Janssen M, Pratt SC, Fewell JH. 2016. The foundress's dilemma: Group selection for cooperation among queens of the harvester ant, Pogonomyrmex californicus. *Sci. Rep.* [Internet] 6:1–9. Available from: www.nature.com/scientificreports
Shin J-H, Blay S, Mcneney B, Graham J. 2006. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria between Single Nucleotide Polymorphisms. *JSS J. Stat. Softw.* 16.
Shoemaker DDW, Ahrens ME, Ross KG. 2006. Molecular phylogeny of fire ants of the Solenopsis saevissima species-group based on mtDNA sequences. *Mol. Phylogenet. Evol.* 38.
Sirviö A, Pamilo P, Johnson RA, Page RE, Gadau J. 2011. Origin and evolution of the dependent lineages in the genetic caste determination system of pogonomyrmex ants. *Evolution (N. Y).* [Internet] 65:869–884. Available from: http://doi.wiley.com/10.1111/j.1558-5646.2010.01170.x
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.
Tang K, Thornton KR, Stoneking M. 2007. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome.Wolfe KH, editor. *PLoS Biol.* [Internet] 5:e171. Available from: https://dx.plos.org/10.1371/journal.pbio.0050171
Tollefson GA, Schuster J, Gelin F, Agudelo A, Ragavendran A, Restrepo I, Stey P, Padbury J, Uzun A. 2019. VIVA (VIsualization of VAriants): A VCF File Visualization Tool. *Sci. Rep.* [Internet] 9:1–7. Available from: https://doi.org/10.1038/s41598-019-49114-z
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:0446–0458.
Wang J, Wurm Y, Nipitwattanaphon M, Riba-Grognuz O, Huang YC, Shoemaker D, Keller L. 2013. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* 493:664–668.
Wilson EO. 1971. The insects societies.
Yan Z, Martin SH, Gotzek D, Arsenault S V., Duchen P, Helleu Q, Riba-Grognuz O, Hunt BG, Salamin N, Shoemaker DW, et al. 2020. Evolution of a supergene that regulates a trans-species social polymorphism. *Nat. Ecol. Evol.* [Internet] 4:240–249. Available from: http://dx.doi.org/10.1038/s41559-019-1081-1
Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen JQ, Hurst LD, Tian D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. *Nature* 523:463–467.
Zeng X, Sun M, Liu L, Chen F, Wei L, Xie W. 2007. Neurexin-1 is required for synapse formation and larvae associative learning in Drosophila. *FEBS Lett.* [Internet] 581:2509–2516. Available from: https://febs.onlinelibrary.wiley.com/doi/full/10.1016/j.febslet.2007.04.068

941 **Data Accessibility**

942 The genome assembly and raw sequencing data prior to trimming and mapping are available at NCBI
943 (BioProject: PRJNA682388; to be released). A detailed description of our bioinformatic analysis pipelines
944 will be made public on https://github.com/schraderL and https://github.com/merrbii/.

945

946 **Author Contributions**

947 JG, UE, ME and LS designed the research, AL, LS, ME, UE and JG analyzed data, ME, UE, LS and JG wrote
948 the manuscript, JG and LS coordinated, all authors revised the manuscript.

949

950 **Figure legends**

951 **Figure 1**: **Population structure and demographic history in *P. californicus* populations.** (A) Principal
952 Component Analysis based on genome-wide SNP data of queens of the pleometrotic (green) and
953 haplometrotic (blue) populations. (B) fineSTRUCTURE coancestry heatmap showing the number of shared
954 haplotypes between donor (columns) and recipient (rows) between individual queen pairs. ADMIXTURE
955 plot below the heatmap shows patterns of ancestry among populations of *P. californicus* at k=2 (additional
956 k values can be found in Supporting Figure S5). Note that one queen (P99P) is an apparent hybrid between
957 the two populations. (C) Demographic history in *P. californicus* computed using MSMC2 with eight phased
958 haplotypes (i.e. four queens), representing the pleometrotic and haplometrotic populations. Effective
959 population size (N_e) is displayed in the top panel and the relative cross-coalescence rate (rCCR) in the lower
960 panel. Lines and shaded areas are means and 95% confidence intervals, respectively.

961

962 **Figure 2**: **Higher Inbreeding and Linkage (and lower heterozygosity) in the pleometrotic population.** (A)
963 Nucleotide diversity (π), (B) heterozygosity, (C) inbreeding coefficient and (D) Tajima's *D* in the
964 pleometrotic (green) and haplometrotic (blue) populations. Heterozygosity and inbreeding coefficient
965 were calculated on a per individual basis, while the other estimates were calculated from 100-kb non-
966 overlapping windows across the genome and averaged for each population. (E) Linkage disequilibrium (LD)
967 decay in the two *P. californicus* populations. LD as measured by pairwise r^2 is represented on the y-axis
968 and distance between pairs of markers within 10 kb on the x-axis. Lines and shaded areas are means and
969 95% confidence intervals (very narrow and hence hardly visible), respectively. For details on statistical
970 tests see main text (***: $P < 0.0001$)

971

972 **Figure 3: Genomic architecture and recombination rates at linkage group (LG) 14 harboring a putative**
973 **supergene.** (A) Genotypes for 108 sons from a single monogynous queen (H-population) based on RADseq
974 markers and for the 35 queen samples from the pleometrotic and haplometrotic populations based on
975 sites corresponding to RADseq markers from the whole genome resequencing data (gray bar-queens).
976 Each row shows a SNP marker (182 in total) distributed across seven scaffolds on LG14. Columns (green
977 ticks at the bottom for pleometrotic and blue ticks for haplometrotic) represent individual samples
978 grouped by population (pleometrotic and haplometrotic) and dataset (males and queens). The black
979 rectangle highlights the non-recombining putative supergene region, spanning scaffolds 22, 23, 38, 42, 48
980 and parts of scaffolds 6 and 14. Red arrowhead indicates the hybrid pleometrotic queen (P99P). (B)
981 Principal Component Analysis based on 65 markers spanning the supergene region showing that the 108
982 male haplotypes are either Sh or Sp. Diploid queens on the other hand are either homozygous for the Sp
983 allele (most of pleometrotic queens, green dots) or heterozygous Sh/Sp (most of the haplometrotic
984 queens, blue dots). (C) Sliding window analyses of LG14, showing (from top to bottom) genetic
985 differentiation between populations (F_{st}), nucleotide diversity (π) and Tajima's *D* estimates in the
986 pleometrotic (green) and haplometrotic (blue) populations.

987

988 **Figure 4**: **Signatures of selection in non-supergene regions**. (A) Distribution of pairwise genetic
989 differentiation (F_{st}) between haplometrotic and pleometrotic queens, estimated in 100-kb non-
990 overlapping windows across the genome. Regions falling into the upper 5% tail of the distribution (F_{st} >
991 0.423, delimited by the blue dashed line) were considered candidates for divergent selection between
992 populations. (B) Sliding window analyses of scaffolds 1, 9, 15 and 16 showing (from top to bottom) genetic
993 differentiation between populations (F_{st}), nucleotide diversity, Tajima's *D* and the cross-population
994 extended haplotype homozygosity (*xpEHH*). Horizontal blue dashed line indicates F_{st} threshold (F_{st} > 0.423;
995 red dots), while horizontal black dashed line represents threshold of significance for outlier SNPs ($p < 0.05$;
996 red dots) in the *xpEHH* analysis. Five candidate regions in four scaffolds showing signatures of selection in
997 the pleometrotic populations are highlighted in red in all panels. These regions are highly differentiated,
998 have reduced diversity and Tajima's *D* in the pleometrotic population and show a significant *xpEHH* peak.
999 Lines and shaded areas are means and 95% confidence intervals, respectively. Data for other scaffolds (1-
1000 25) are shown in Supporting Figure S6. (C) Signatures of site-specific extended haplotype homozygosity
1001 (*EHHS*) at two outlier SNPs in the pleometrotic (green), but not the haplometrotic population (blue)
1002 indicating recent selective sweeps.

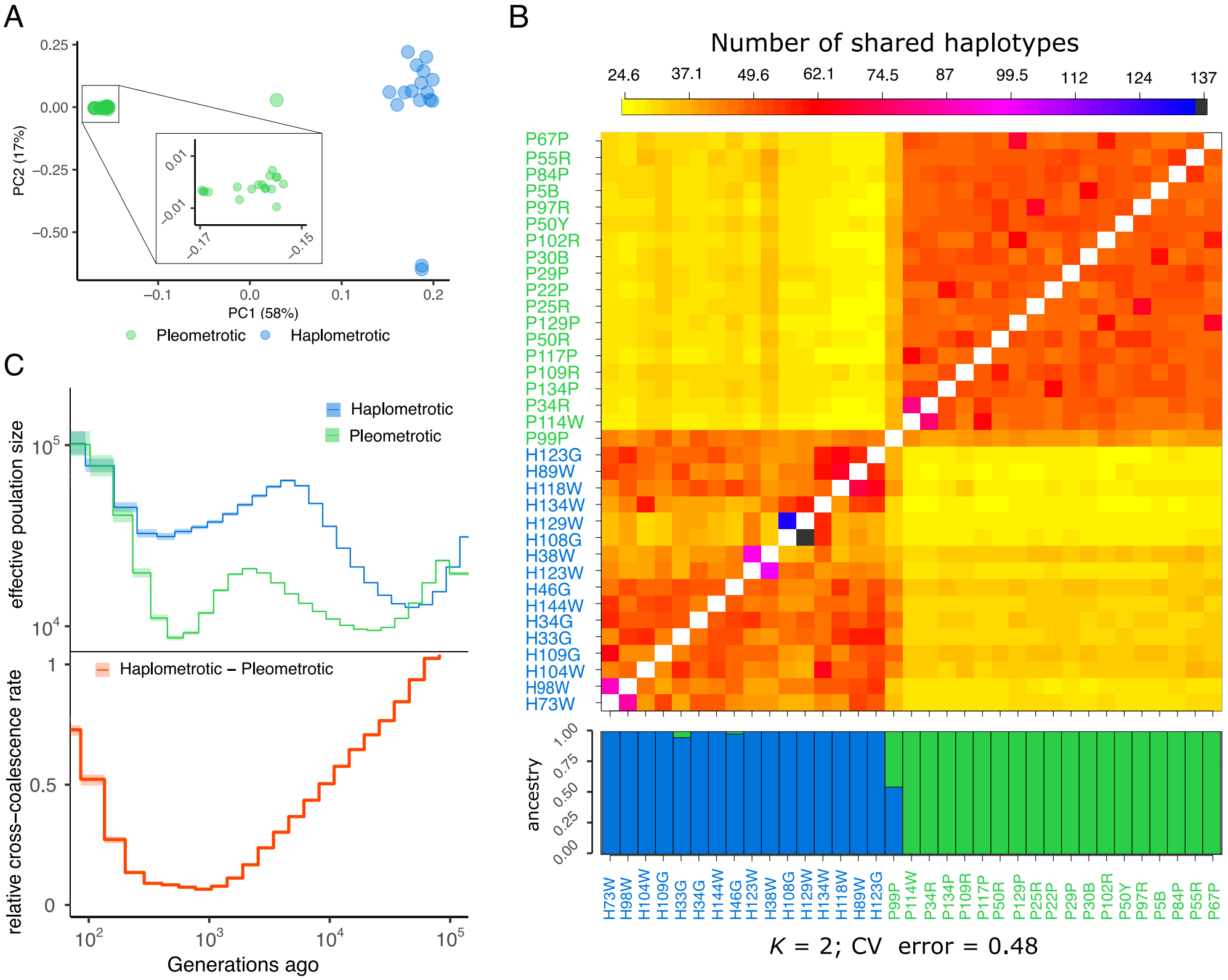
1003

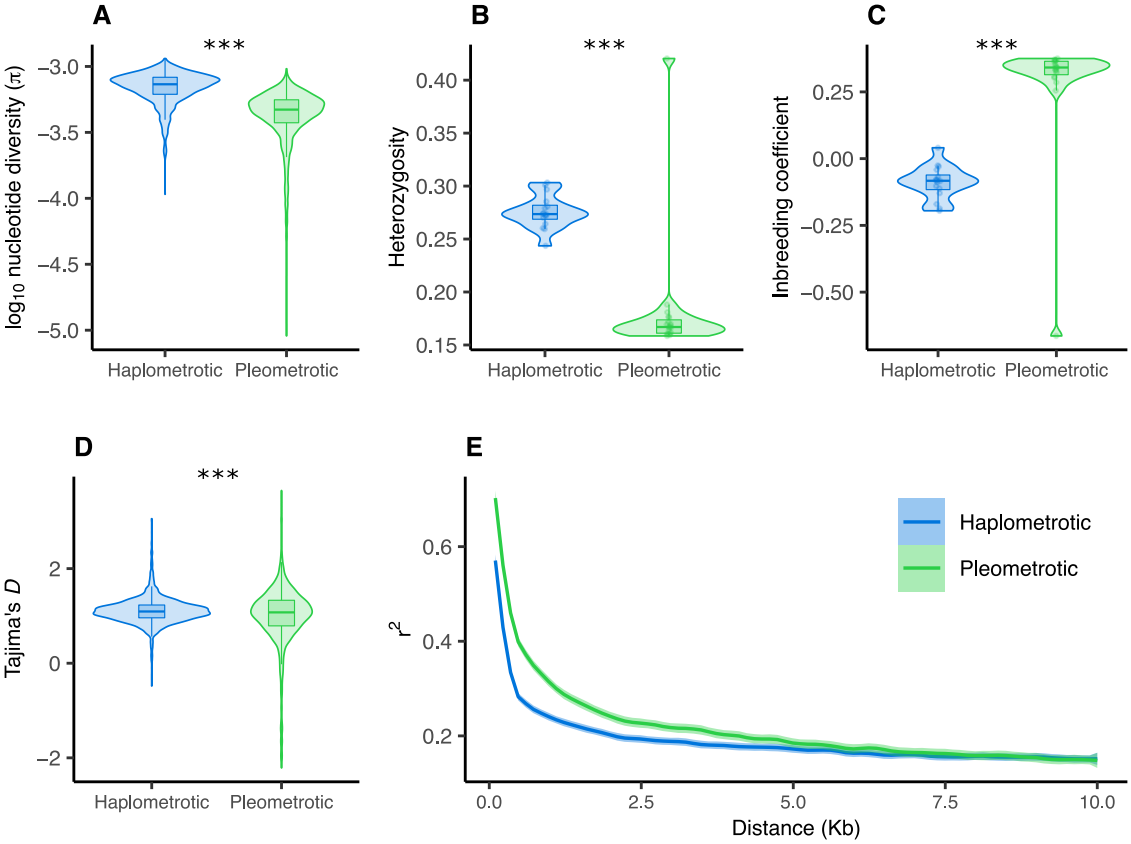
1004 **Table legend:**

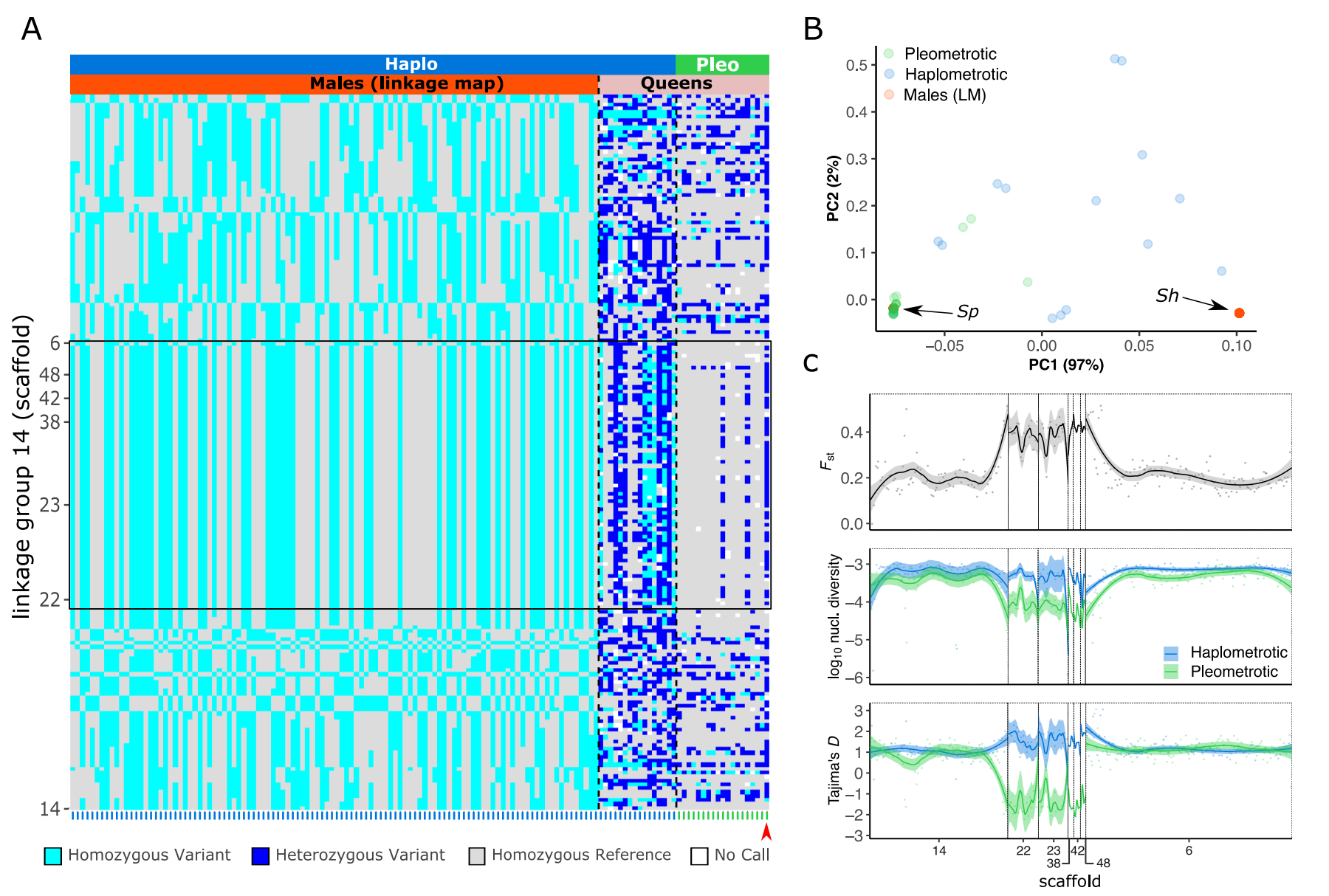
1005

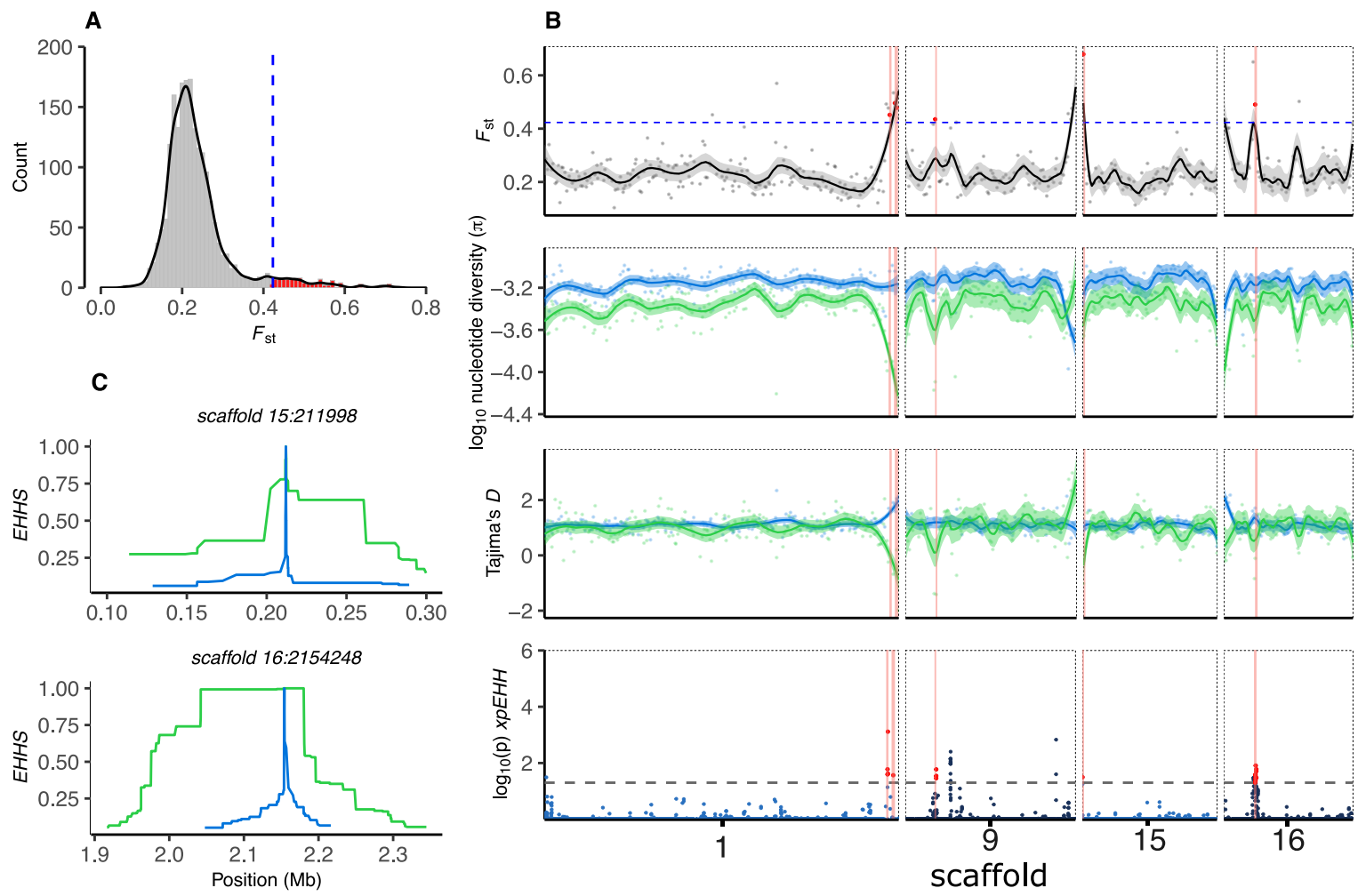
1006 **Table 1:** Effects of single nucleotide polymorphisms (SNPs) found in our five candidate regions that show
1007 strong signatures of selection in the pleometrotic population. We annotated 6,969 putative effects on
1008 protein coding genes of 3,782 SNPs.

1009


A
PC2 (17%)
PC1 (58%)
Pleometrotic
Haplometrotic
B
Number of shared haplotypes
24.6
37.1
49.6
62.1
74.5
87
99.5
112
124
137
ancestry
K = 2; CV error = 0.48
C
Haplometrotic
Pleometrotic
effective poulation size
Haplometrotic – Pleometrotic
relative cross–coalescence rate
Generations ago


A
***
log10 nucleotide diversity (π)
−3.0
−3.5
−4.0
−4.5
−5.0
Haplometrotic
Pleometrotic
B
***
Heterozygosity
0.40
0.35
0.30
0.25
0.20
0.15
Haplometrotic
Pleometrotic
C
***
Inbreeding coefficient
0.25
0.00
−0.25
−0.50
Haplometrotic
Pleometrotic
D
***
Tajima's D
2
0
−2
Haplometrotic
Pleometrotic
E
Haplometrotic
Pleometrotic
r2
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
Distance (Kb)


A
Haplo
Pleo
Males (linkage map)
Queens
linkage group 14 (scaffold)
6
48
42
38
23
22
14
Homozygous Variant
Heterozygous Variant
Homozygous Reference
No Call
B
Pleometrotic
Haplometrotic
Males (LM)
PC2 (2%)
PC1 (97%)
Sp
Sh
C
F_{st}
$\log_{10}$ nucl. diversity
Haplometrotic
Pleometrotic
Tajima's D
scaffold


A
Count
F_{st}
B
F_{st}
$\log_{10}$ nucleotide diversity (π)
Tajima's D
$\log_{10}$(p) $xpEHH$
scaffold
C
scaffold 15:211998
scaffold 16:2154248
EHHS
Position (Mb)

Effect	Count	Proportion [%]
3’UTR	19	0.27
5’UTR	3	0.04
Downstream	1563	22.43
Intergenic	3385	48.57
Intron	209	3.00
Nonsynonymous variant	99	1.42
Synonymous variant	84	1.21
Upstream	1597	22.92
Other	10	0.14

Supporting Figures for:

Genomic architecture and evolutionary dynamics of a social niche polymorphism in the California harvester ant, *Pogonomyrmex californicus*

Running title:

California harvester ant population genomics

Mohammed Errbii[1], Ulrich R. Ernst[1,§], Aparna Lajmi[2], Jürgen Gadau[1,3,‡], Lukas Schrader[1,3,‡]

[1] Molecular Evolution and Sociobiology Group, Institute for Evolution & Biodiversity, University of Münster, Hüfferstr. 1, DE-48149 Münster, Germany

[2] Department of Evolutionary and Environmental Biology, Institute of Evolution, University of Haifa, Haifa, Israel

[3] lukas.schrader@uni-muenster.de; gadauj@uni-muenster.de

[‡] equal contribution

[§] current address: Apicultural State Institute, University of Hohenheim, August-von-Hartmann-Str. 13, DE-70599 Stuttgart, Germany

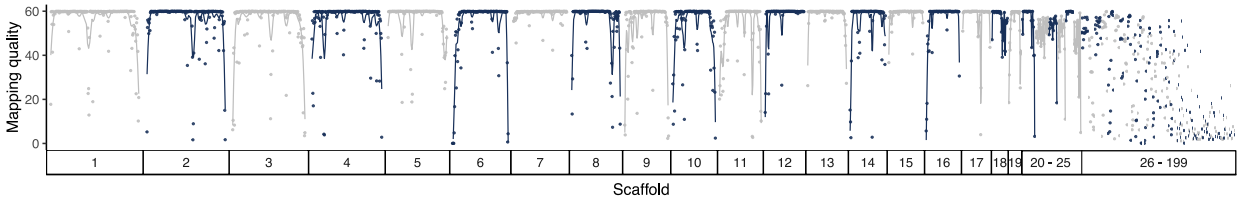


Figure S1: Mapping quality across the 199 scaffolds of the Pcal2 assembly. Dots represent average mapping quality in 100 kb window, calculated and averaged across the 35 queens sequenced in this study. Scaffolds 26-199 were excluded from all analyses due to fragmentation and low mapping quality.

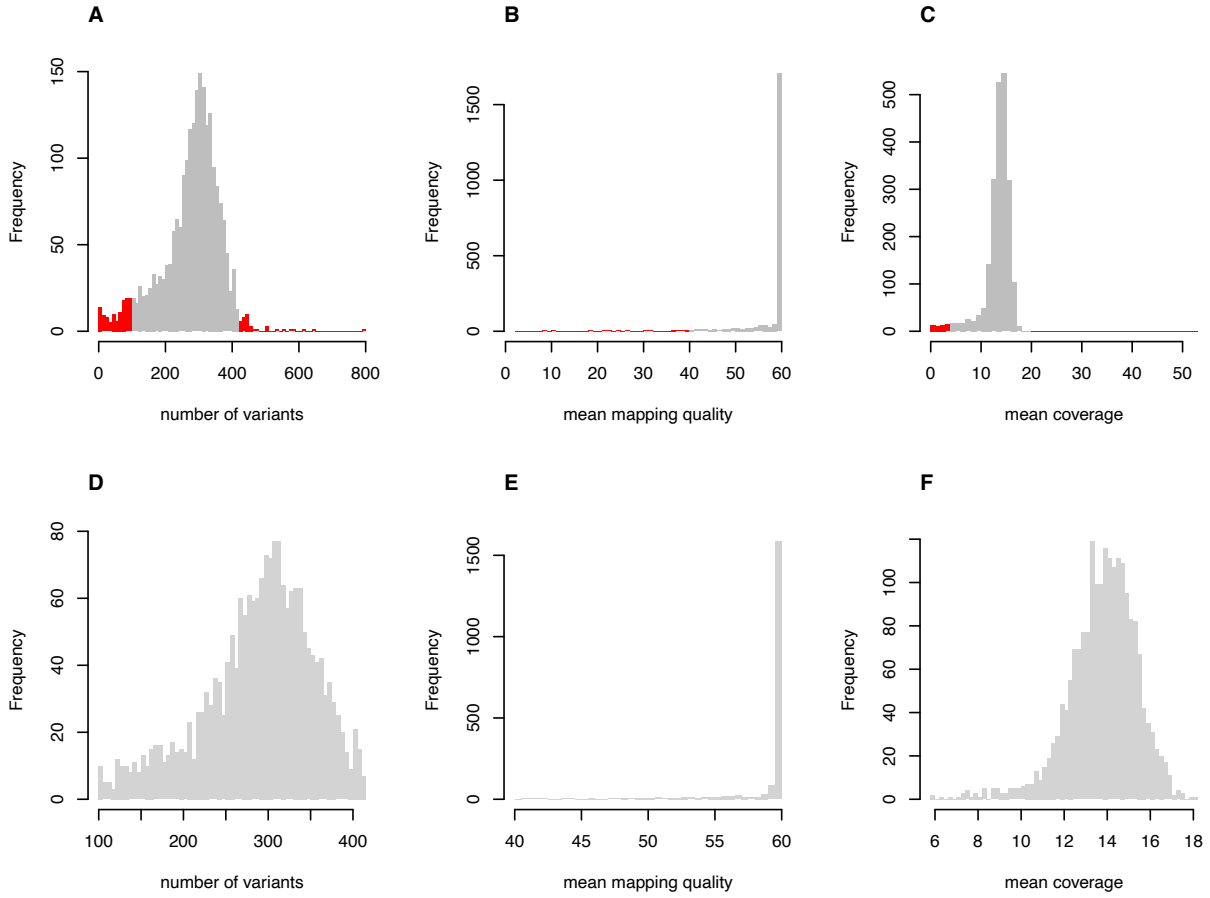


Figure S2: Quality control of non-overlapping genomic windows (100 kb) used to compute population genomic metrics. Number of variants, mapping quality and mean coverage per 100 kb window before (A, B, and C, respectively) and after (D, E, and F, respectively) applying cutoffs (see Material and Methods). Red bars indicate windows excluded from population genomic analyses.

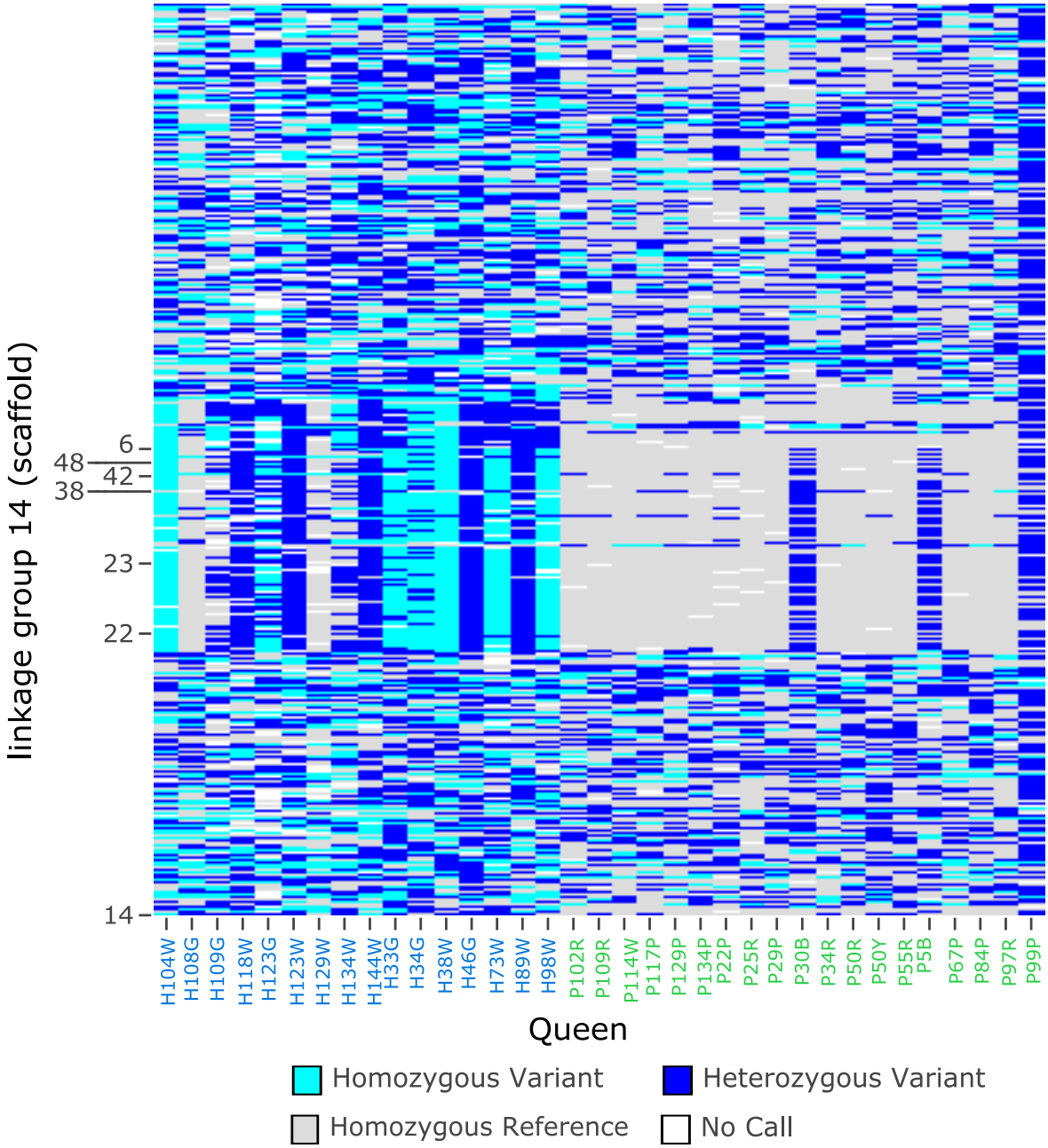


Figure S3: Genomic architecture at linkage group (LG) 14 harboring the supergene in *P. californicus*. The heatmap displays the genotypes of the 35 queen samples from the pleometrotic and haplometrotic populations. Rows represents a total of 5,363 SNPs (thinned out of 68,643 SNPs for ease of visualization) with a MAF ≥ 0.2, spanning seven scaffolds of LG14. Columns represent individual queens grouped by population (pleometrotic and haplometrotic). Black rectangle delimits the non-recombining supergene region, spanning scaffolds 22, 23, 38, 42, 48 and parts of scaffolds 6 and 14.

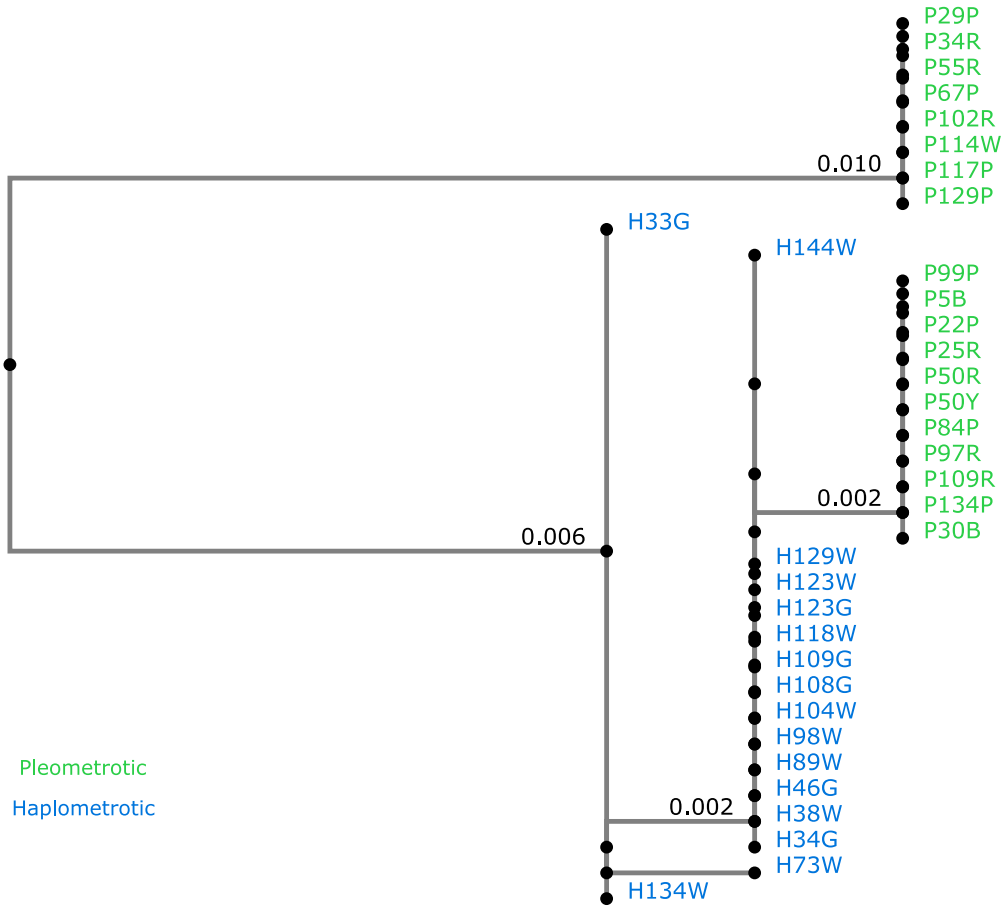


Figure S4: Neighbor-joining tree based on partial *cytochrome oxidase I* (COI) sequences of all 35 *P. californicus* queens sequenced in this study. Numbers displayed on branches show branch length. The pleometrotic population had only two haplotypes, that were all differentiated from the 3 haplotypes of the haplometrotic population.

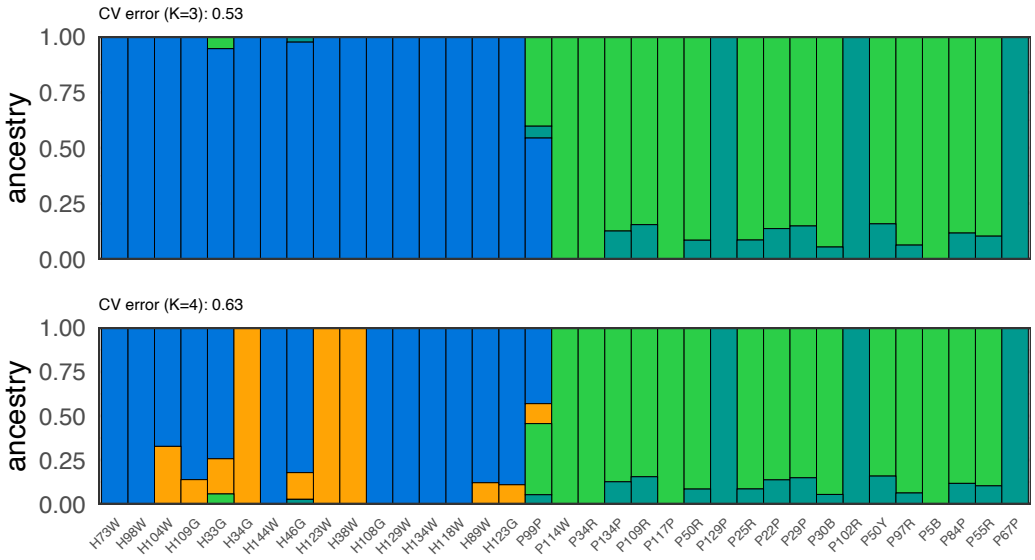


Figure S5: Admixture patterns among *P. californicus* populations by using ADMIXTURE at K = 3 and K = 4. For each value of k, the cross-validation error rate is presented.

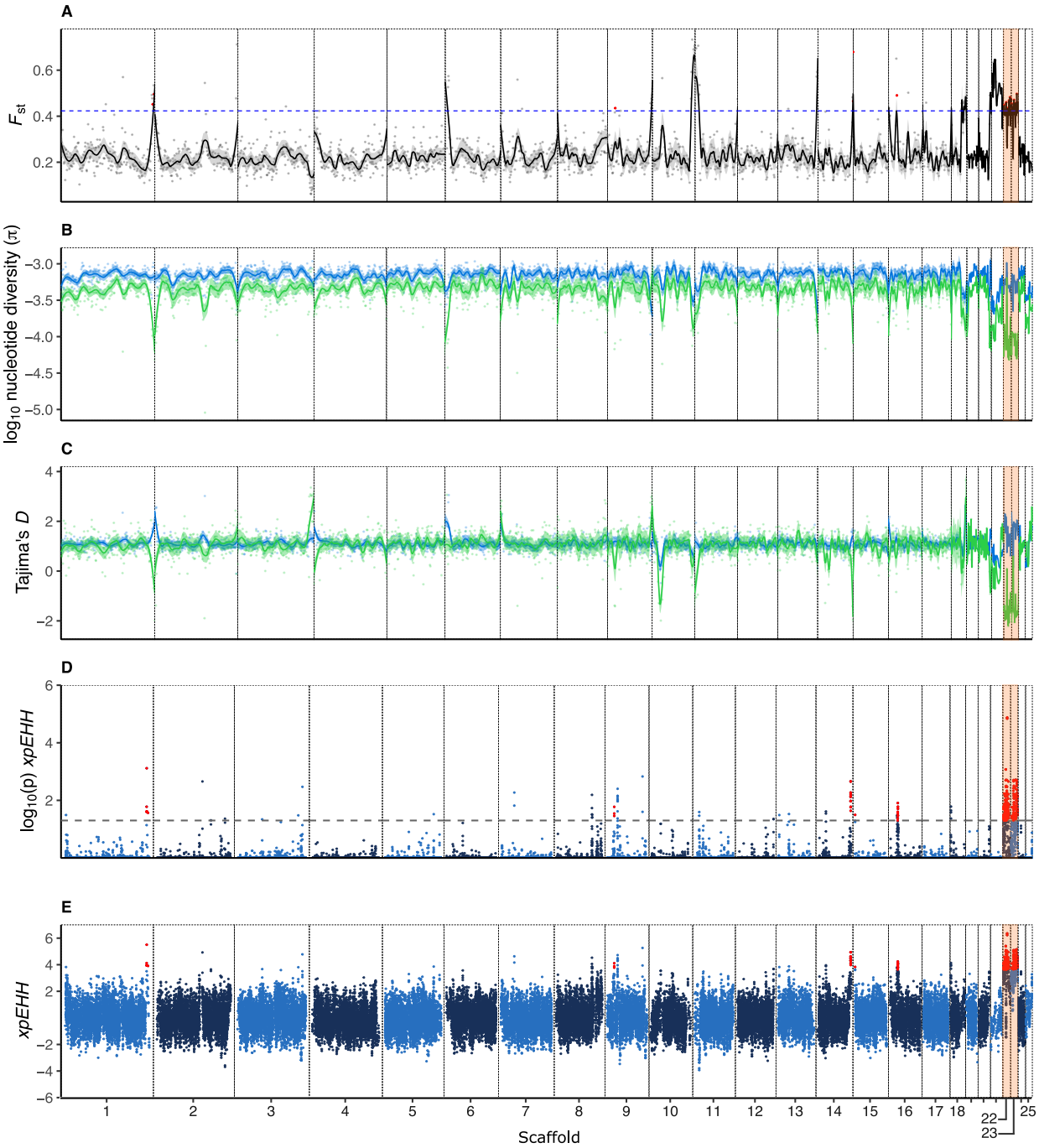


Figure S6: Genome-wide distribution of genetic differentiation (F_{st}), nucleotide diversity (π), Tajima's *D* and *xpEHH* scores in the 25 largest scaffolds of the *P. californicus* genome. (A) Genetic differentiation (F_{st}) between haplometrotic and pleometrotic queens. Blue dashed line indicates the 5% cutoff (cf. Fig. 4A). (B) Nucleotide diversity (π) and (C) Tajima's *D* in the haplometrotic (blue) and pleometrotic (green) populations. All estimates were calculated in 100 kb non-overlapping windows. Lines and shaded areas are means and 95% confidence intervals, respectively. (D) log-scaled P-values and (E) *xpEHH* scores associated with SNPs across the 25 scaffolds. Red dots highlight windows (F_{st}) and SNPs (*xpEHH*) showing signatures of selection in both analyses. Scaffolds 22 and 23 containing the putative supergene are highlighted in orange.

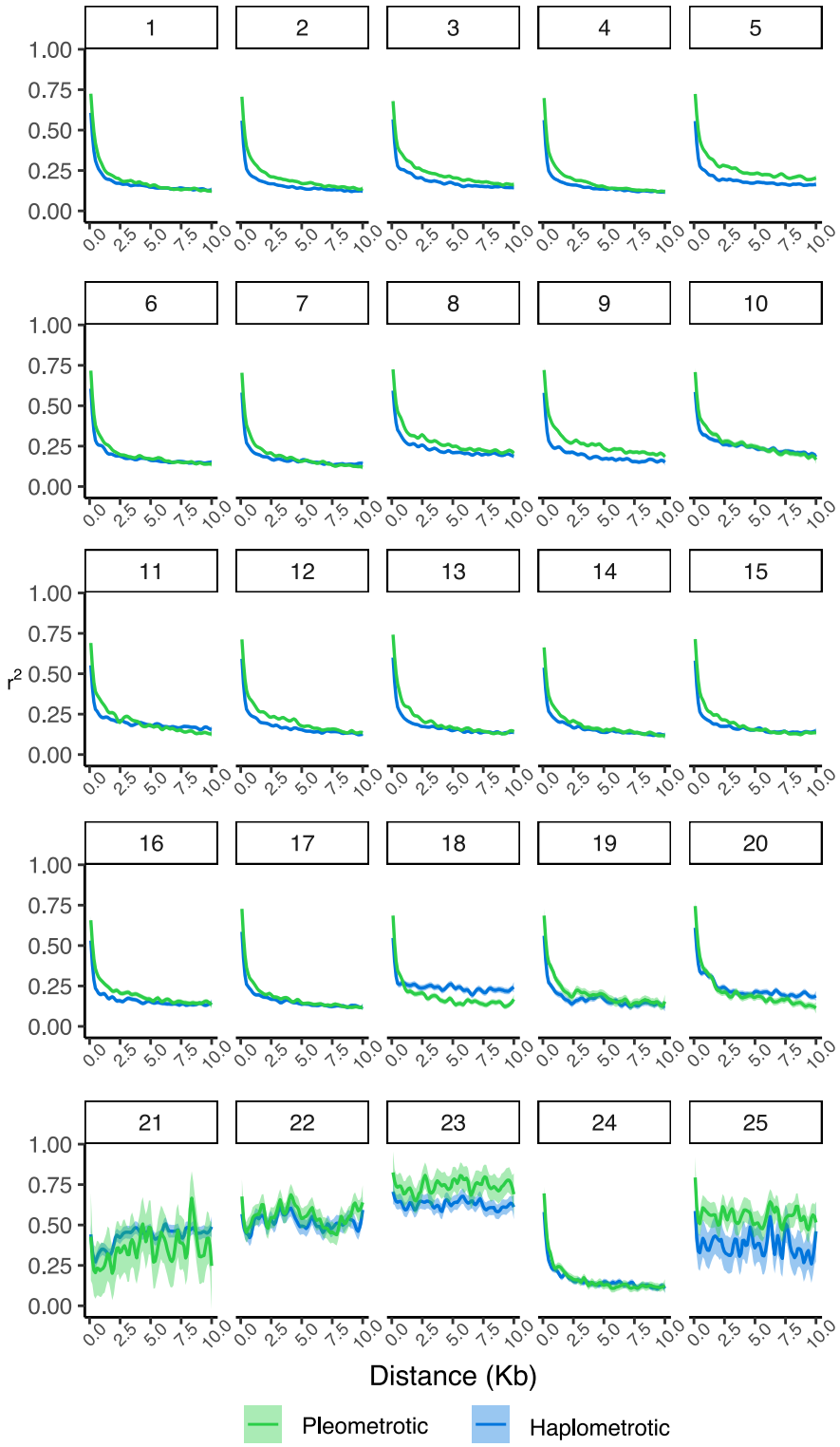


Figure S7: Linkage disequilibrium (LD) decay displayed for each of the 25 largest scaffolds in the two *P. californicus* populations. LD as measured by pairwise r^2 is represented on the y-axis and distance between pairs of markers within 10 kb on the x-axis. Lines and shaded areas are means and 95% confidence intervals, respectively.

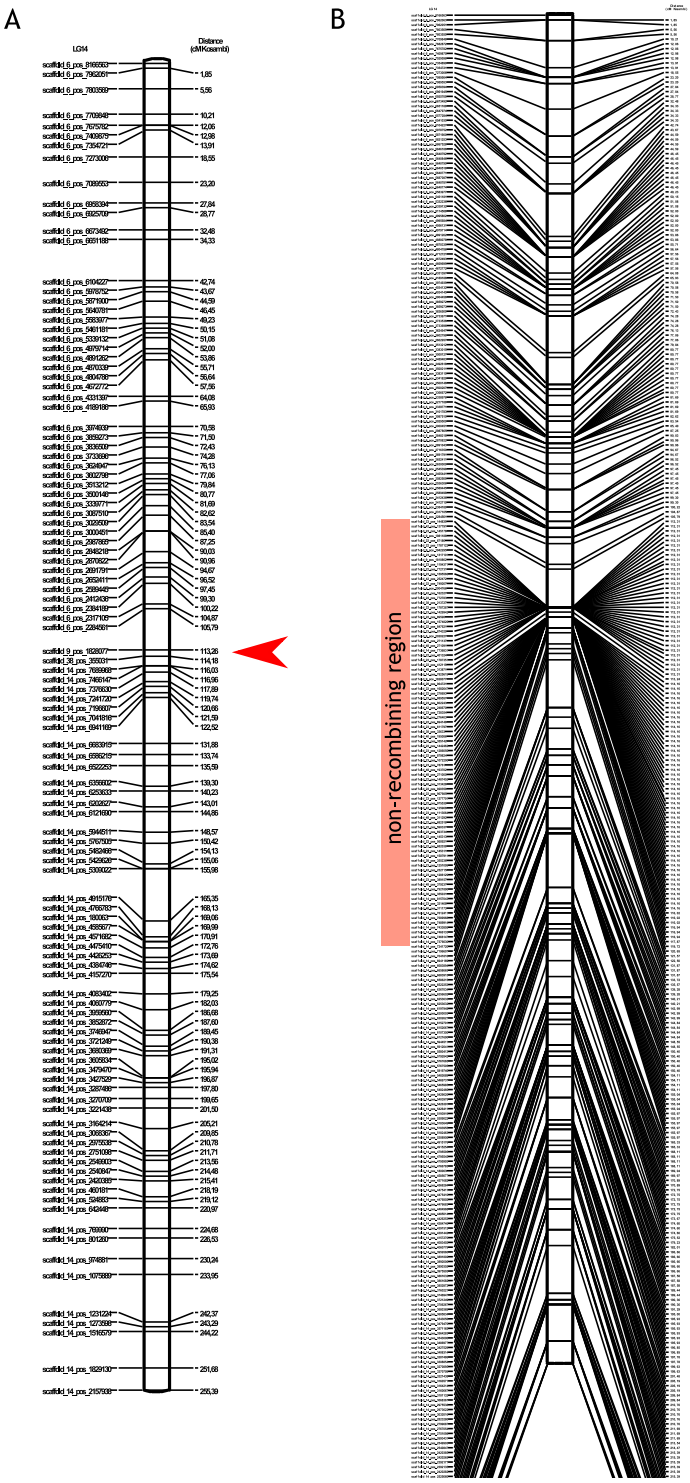


Figure S8: Linkage group (LG) 14 (253.38 cM). Marker names are given on the left and genetic distance in cM (Kosambi) are given on the right sides. Both figures represent the same LG group, panel B includes all 339 markers and shows the non-recombining area. Panel A shows a "skeleton linkage group" where overlapping markers at the same position are grouped and represented only by a single marker name (i.e. the non-recombining region is represented by a single marker; thick red arrowhead).

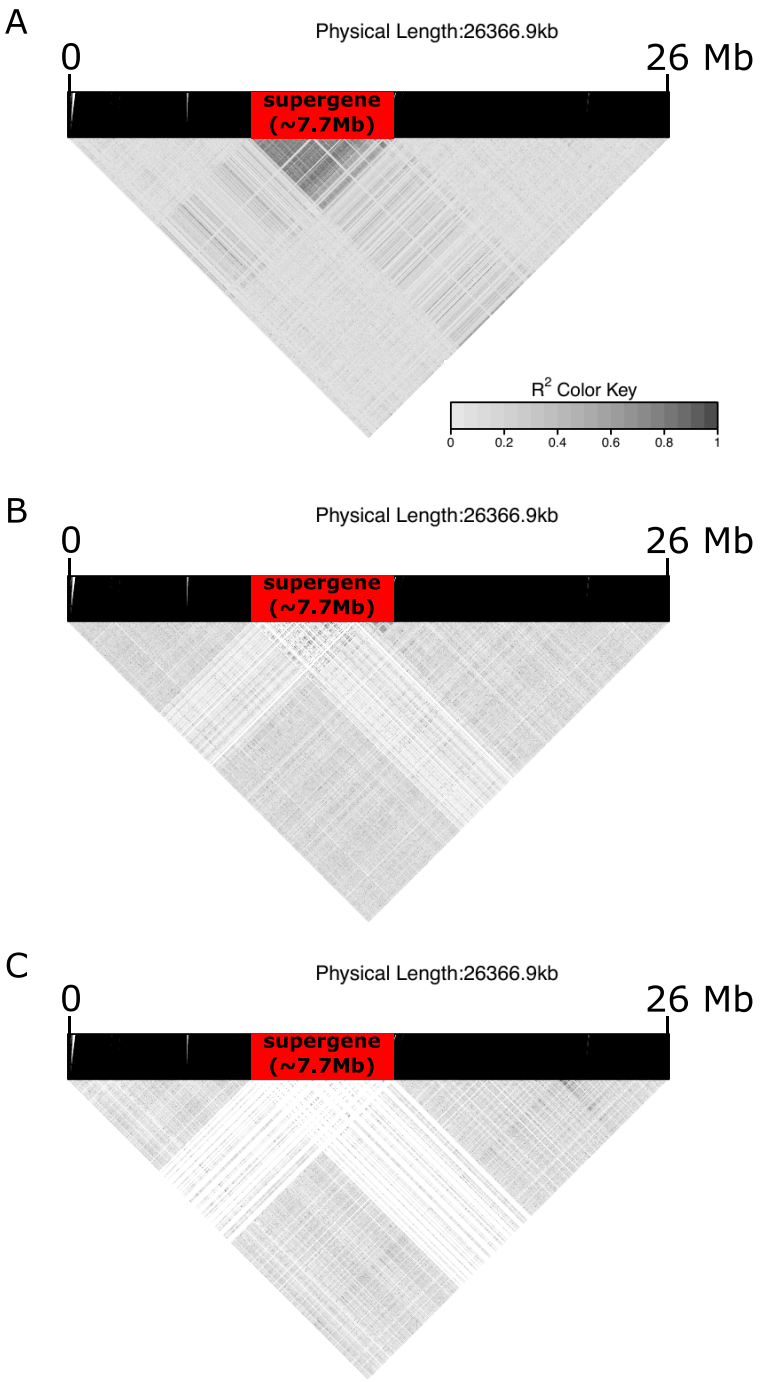


Figure S9: Linkage disequilibrium estimates (r^2) across linkage group (LG) 14 containing the supergene. (A) Pairwise LD for pooled queens homozygous for Sh (n=6) and Sp (n=6). (B) Pairwise LD for Sh homozygous queens and (C) Sp homozygous queens only. The red rectangle shows the region of no recombination between the Sh and Sp haplotypes. SNPs are ordered according to physical position on the chromosome after lift over. Black lines on top of each heatmap indicate SNP positions on the physical map. Note that invariant marker sites are displayed in white, dominating particularly the homozygous Sh/Sh and Sp/Sp LD plots.

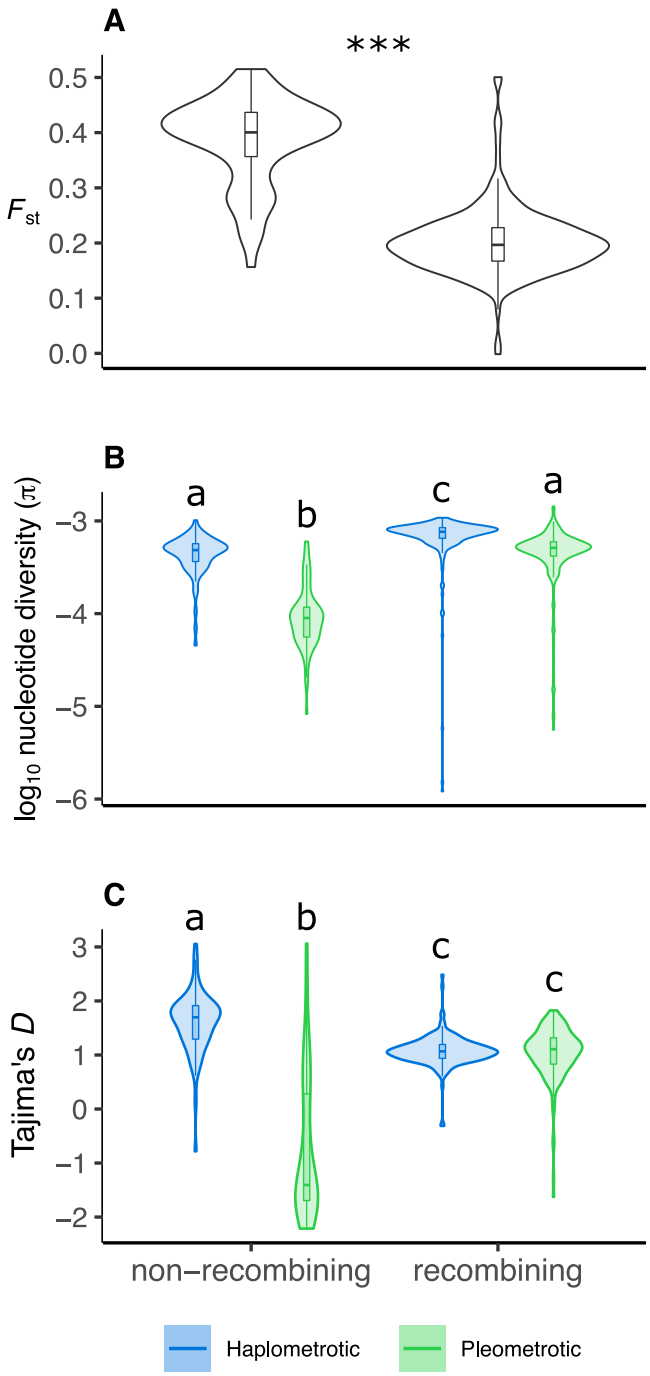


Figure S10: Genetic differentiation and diversity within recombining and non-recombining regions of LG14. (A) Genetic differentiation (F_{st}), (B) nucleotide diversity (π) and (D) Tajima's *D* in the pleometrotic and haplometrotic populations of *P. californicus*. For F_{st}, we performed a Wilcoxon rank sum test (W = 12261, p < 2.2e-16). We performed a Kruskal-Wallis rank sum test for π (χ^2 = 254.77, df = 3, p < 2.2e-16) and Tajima's *D* (χ^2 = 177.81, df = 3, p < 2.2e-16), followed by pairwise Wilcoxon rank sum *post hoc* tests. Different letters represent significant differences (p < 0.0001).

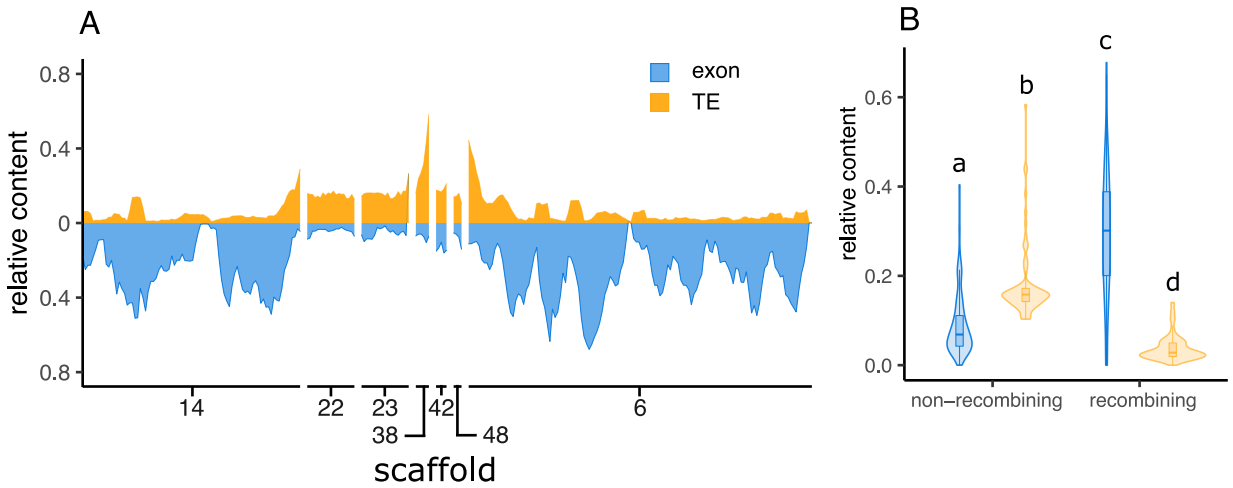


Figure S11: Reduced exon and elevated TE content in the supergene region on LG14 (A) Genomic distribution of exon and TE sequence along LG14. (B) Violin plots showing exon and TE content within non-recombining regions (i.e. the supergene) and recombining regions of LG14. We performed a Kruskal-Wallis rank sum test (χ^2 = 342.74, df = 3, p < 2.2e-16), followed by pairwise Wilcoxon rank sum *post hoc* tests. Different letters represent significant differences (p < 0.0001).

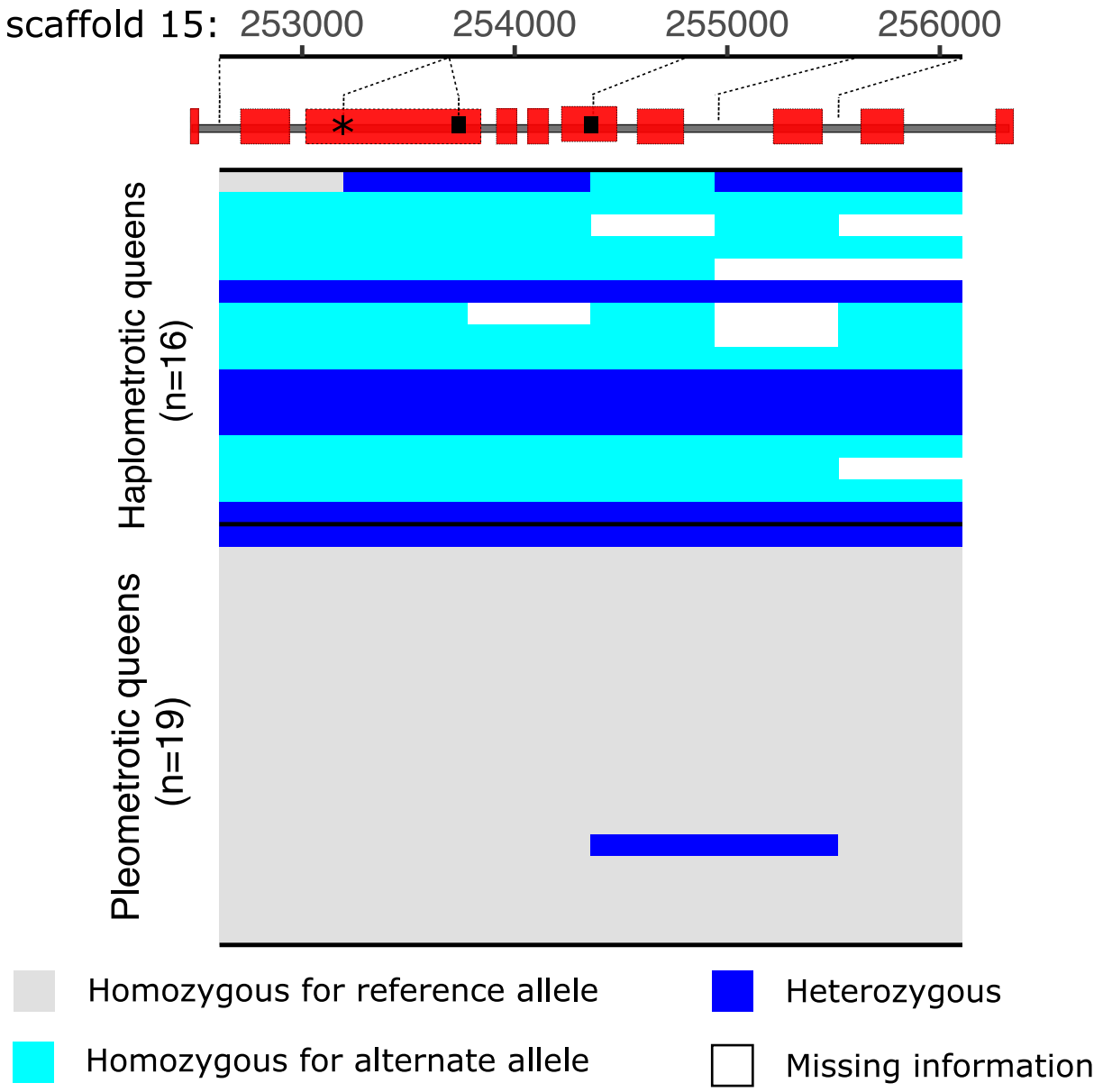


Figure S12: Genotypes of haplo- and pleometrotic queens for one candidate gene under selection. Plot showing six SNP positions (top, dashed lines), the affected features of the *DNMT1* (*DNA methyltransferase 1*) gene, and the genotypes for each queen in the two populations. (*) indicates a non-synonymous substitution and (■) indicates a synonymous substitution. Exons are shown by red rectangles. Pleometrotic queens are almost all homozygous and fixed for the reference allele, unlike the haplometrotic queens that are more variable. Plot was produced using GenotypePlot [https://github.com/JimWhiting91/genotype_plot]. Note that the top row in the pleometrotic population displays the genotype of the hybrid P99P queen.

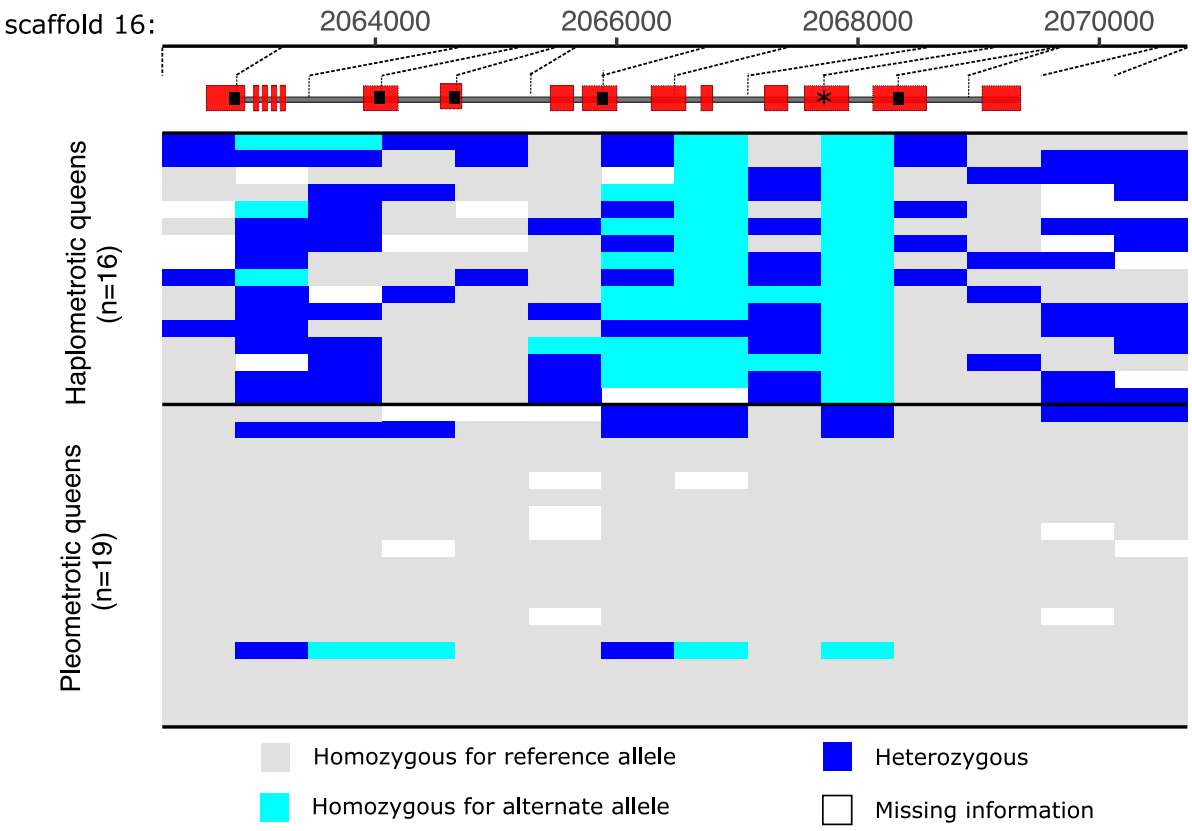


Figure S13: Genotype distribution between haplo- and pleometrotic queens for one candidate gene Plot showing 13 SNP positions (top, dashed lines), the affected features of the *Histone acetyltransferase MYST2* gene, and the genotype of each queen in the two populations. (*) indicates a non-synonymous substitution and (■) indicates a synonymous substitution. Exons are shown by red rectangles. Pleometrotic queens are almost all homozygous and fixed for the reference allele, unlike the haplometrotic queens that are more variable. Plot was produced using GenotypePlot [https://github.com/JimWhiting91/genotype_plot].