# Evolution of a supergene that regulates a trans-species social polymorphism

Zheng Yan, Simon H Martin, Dietrich Gotzek, Samuel V Arsenault, Pablo Duchen, Quentin Helleu, Oksana Riba-Grognuz, Brendan G Hunt, Nicolas Salamin, Dewayne Shoemaker, et al.

nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-019-1081-1

# Evolution of a supergene that regulates a trans-species social polymorphism

Zheng Yan[1], Simon H. Martin[2], Dietrich Gotzek[3], Samuel V. Arsenault[4], Pablo Duchen[5], Quentin Helleu[1], Oksana Riba-Grognuz[1], Brendan G. Hunt[4], Nicolas Salamin[5], DeWayne Shoemaker[6], Kenneth G. Ross[4*] and Laurent Keller[1*]

**Supergenes are clusters of linked genetic loci that jointly affect the expression of complex phenotypes, such as social organization. Little is known about the origin and evolution of these intriguing genomic elements. Here we analyse whole-genome sequences of males from native populations of six fire ant species and show that variation in social organization is under the control of a novel supergene haplotype (termed *Sb*), which evolved by sequential incorporation of three inversions spanning half of a 'social chromosome'. Two of the inversions interrupt protein-coding genes, resulting in the increased expression of one gene and modest truncation in the primary protein structure of another. All six socially polymorphic species studied harbour the same three inversions, with the single origin of the supergene in their common ancestor inferred by phylogenomic analyses to have occurred half a million years ago. The persistence of *Sb* along with the ancestral *SB* haplotype through multiple speciation events provides a striking example of a functionally important trans-species social polymorphism presumably maintained by balancing selection. We found that while recombination between the *Sb* and *SB* haplotypes is severely restricted in all species, a low level of gene flux between the haplotypes has occurred following the appearance of the inversions, potentially mitigating the evolutionary degeneration expected at genomic regions that cannot freely recombine. These results provide a detailed picture of the structural genomic innovations involved in the formation of a supergene controlling a complex social phenotype.**

While it is becoming increasingly clear that many animal species exhibit variation in social organization, the underlying causes are rarely understood. The first discovery of a genetic basis for such variation was in the fire ant *Solenopsis invicta*[1,2]. In this species, variation at a genomic region containing the odorant-binding protein gene *Gp-9* determines whether colonies contain just one (monogyne social form) or multiple (polygyne form) queens[2,3], a fundamental distinction associated with a suite of other important individual- and colony-level phenotypic differences[4]. Studies of invasive US populations revealed that *Gp-9* is located in a supergene on the 'social chromosome' (chromosome 16) and that the *Social b* (*Sb*) haplotype harbouring the $Gp\text{-}9^{b}$ allele apparently does not recombine with the *Social B* (*SB*) haplotype containing the alternate $Gp\text{-}9^{B}$ allele[5]. In the United States, monogyne colonies invariably contain a single homozygous *SB/SB* queen and only *SB/SB* workers, while polygyne colonies always contain multiple heterozygous (*SB/Sb*) queens together with predominantly *SB/Sb* and *SB/SB* workers (*Sb/Sb* females have low viability and *SB/SB* queens are killed by nestmate workers in polygyne colonies[4,6]).

Reconstruction of the routes of supergene evolution is a long-standing goal with important implications for our understanding of how these remarkable genomic entities come to regulate the myriad features of complex phenotypes[7–10]. We conducted a comparative genomic study of several fire ant species sampled from their native ranges to characterize fully variation at the fire ant supergene and, thereby, explain its origin and subsequent evolution.

## Results and discussion

A previous study showed that the supergene in *S. invicta* contains two inversions, with a third one suggested by the fact that strong linkage extends approximately 1 megabase (Mb) beyond the two identified inversions[11]. To test this prediction and characterize all of the inversion breakpoints, we assembled new genomes of two *S. invicta* males (ant males are haploid), one *Sb* and one *SB* male from the same polygyne colony from the United States, using long (>10 kilobases, kb) PacBio sequence reads. Pairwise alignment of these two newly assembled genomes revealed no sign of structural rearrangements at 15 of the chromosomes (Extended Data Fig. 1) but did reveal three large inversions in the *Sb* haplotype which, collectively, span ~11.4 Mb of chromosome 16 (Fig. 1a).

The largest inversion, *In(16)1* (9.48 Mb), contains 476 annotated protein-coding genes (Supplementary Table 1), including *Gp-9*. The proximal breakpoint disrupts the '*F-box/WD repeat-containing protein 4-like*' gene (*FBXW4*; LOC105199310) six nucleotides downstream from its start codon and 26 nucleotides downstream from the *SB* transcript start site. The distal breakpoint disrupts the '*Phosphoglycerate mutase 2*' gene (*PGAM2*; LOC105193833) eight nucleotides downstream from the *SB* transcript start site in its 5′ untranslated region, UTR (Fig. 1b and Extended Data Fig. 2a–c). Remarkably, comparative RNA-seq analyses show that neither of these genes exhibits consistent differential expression between adults (males, queens and workers) with alternate supergene haplotypes/genotypes (all false discovery rates, FDR, $P > 0.05$ except for one sample type in one gene; Fig. 2), suggesting minimal position

[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland. [2]Institute of Evolutionary Biology, the University of Edinburgh, Edinburgh, UK. [3]Department of Entomology and Laboratories of Analytical Biology, National Museum of Natural History, Smithsonian Institution, Washington, DC, USA. [4]Department of Entomology, University of Georgia, Athens, GA, USA. [5]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [6]Department of Entomology and Plant Pathology, University of Tennessee, Knoxville, TN, USA.
*e-mail: kenross@uga.edu; Laurent.keller@unil.ch

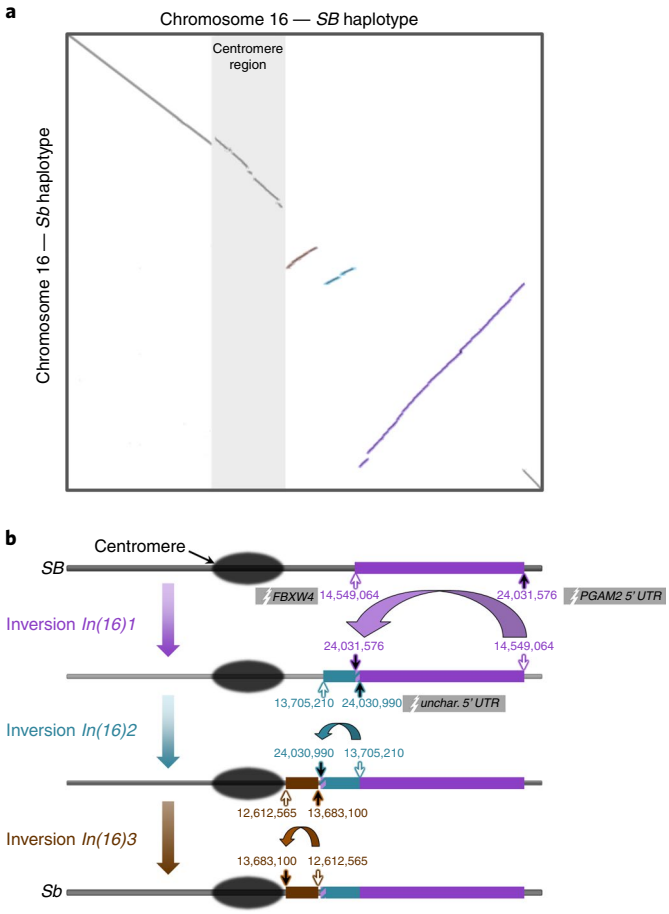



**Fig. 1 | The three inversions that distinguish the *Sb* and *SB* haplotypes on the fire ant social chromosome (chr16). a**, Dot plot of the sequence alignment between chr16 of *Sb* and *SB* males of *S. invicta* (pairwise alignment of NCBI genome ID *Solenopsis_invicta_M01_SB* with *Solenopsis_invicta_M02_Sb*). Grey dots indicate the forward strand alignment and dots of other colours indicate the reverse strand alignment; the three major non-grey lines correspond to the three inversions. The centromere region is defined by the presence of centromeric satellite sequences. **b**, Evolution of the *Sb* haplotype from an *SB*-like haplotype by acquisition of three inversions—*In(16)1, In(16)2* and *In(16)3*—arranged in order of their inferred evolutionary appearance (Extended Data Fig. 9). White and black arrows indicate locations of proximal and distal breakpoints, respectively, of each inversion (with associated genomic coordinates). Note the short overlap (586 nucleotides, nt) between the proximal breakpoint of inversion *In(16)1* and distal breakpoint of *In(16)2*, indicating that this fragment was inverted twice. Coding genes disrupted by the inversions are shown in grey boxes (*FBXW4*: *F-box/WD repeat-containing protein 4-like* gene; *PGAM2 5′ UTR*: *Phosphoglycerate mutase 2* gene 5′ UTR; *unchar. 5′ UTR*: uncharacterized gene 5′ UTR). Structures are not all drawn to scale.

effects of *In(16)1* in the regulation of proximal loci. In the case of *FBXW4*, which probably plays a role in protein degradation[12] (Extended Data Fig. 3), the *Sb* allele retains only the second in-frame start codon of the *SB* transcript, which is seven codons downstream of the first annotated *SB* start codon. Modest truncation of the *Sb* protein product, through use of this alternative translation start site, could feasibly affect its function. However, *SB* translation initiation at the first *FBXW4* start codon has yet to be demonstrated, so different-size proteins encoded by the alternate alleles are not assured.

The second inversion, *In(16)2* (0.84 Mb), contains 46 annotated protein-coding genes (Supplementary Table 2) and overlaps slightly with the first inversion, leading to a short (586 nucleotides) doubly inverted fragment (Fig. 1b and Extended Data Fig. 2d). Notably, homologous copies of a '*Jockey-like mobile element*' are present at each breakpoint of inversion *In(16)2*, suggesting that activity of this element promoted the appearance of the inversion via ectopic (non-allelic homologous) recombination[13] (Extended Data Fig. 2e). The distal breakpoint of this inversion disrupts an uncharacterized gene (LOC105193832) 75 nucleotides downstream of the *SB* transcript start site in its 5′ UTR (Extended Data Fig. 2f). This uncharacterized locus exhibits significant differential expression between individuals with alternate supergene haplotypes/genotypes in all three castes according to RNA-seq analyses (FDR $P < 0.05$, Fig. 2), indicating a widespread position effect of *In(16)2* on transcription of LOC105193832. The near-constitutive nature of this effect is reinforced by consistently higher expression in adults bearing the *Sb* haplotype for males (whole bodies), queens (one of two tissues) and workers (both tissues) (Fig. 2). Thus, somewhat surprisingly given that the breakpoint interrupts the first exon of this gene, the effect is not degenerative; instead, *In(16)2* generally enhances the gene's transcription. This could be explained by the presence of different promoter sequences at LOC105193832 in the *SB* and *Sb* haplotypes or by differences in messenger RNA (mRNA) stability arising from their distinct 5′ UTR sequences. In further support of this proposed upregulation of the *Sb* haplotype at LOC105193832, we found significantly elevated *Sb* haplotype-specific expression at single nucleotide polymorphisms (SNPs) diagnostic for the alternate *S. invicta* haplotypes in *SB/Sb* workers and queens (Supplementary Table 3). These striking patterns raise the possibility that *In(16)2* directly influences trait variation relevant to social organization by altering LOC105193832 expression patterns[11].

The third inversion, *In(16)3* (1.07 Mb), contains 27 protein-coding genes (Supplementary Table 4), none of which is interrupted. The inversion does, however, bridge the region between the centromere (enriched with satellite DNA repeats[14]) and the two other inversions (Fig. 1b), thereby considerably expanding the region of restricted recombination[15] (centromeres typically exhibit reduced recombination[8,14,16]). The two breakpoints of *In(16)3* are in regions with similar tandem repeat structures (Extended Data Fig. 2g–i), suggesting that this inversion also originated by ectopic recombination[13].

The results above expand on the recent findings of Huang et al.[11] in several important respects. First, we confirmed the predicted third supergene inversion (*In(16)3*), the existence of which expands the conceived supergene boundaries into an enlarged region of suppressed recombination (see Fig. 1b and linkage disequilibrium, LD, results below). Second, we were able to generate upgraded inventories of the genes located within each inversion (Supplementary Tables 1,2 and 4), made possible by our improved genome assemblies. Third, we determined the location of the proximal breakpoint of *In(16)1* within the gene *FBXW4*, which putatively causes alternative translation start site usage in the *SB* and *Sb* haplotypes. Fourth, we showed that the homologous copies of a '*Jockey-like mobile element*' present at each breakpoint of inversion *In(16)2* provide a probable mechanistic explanation for its origin. Finally, we greatly extended previous work on the effects of chromosome breakpoints on protein-coding gene expression[11] by analysing distinct castes and tissue types (Fig. 2). The fire ant supergene is known to influence a vast spectrum of individual- and colony-level trait variation involving all castes (mostly in the adult stages) and comprising diverse social and reproductive contexts;[4,17,18] thus, analyses of regulatory effects of the *Sb* haplotype in varied biological settings are necessary to begin to link gene function to the complex trait variation observed in this system.

Examination of new genome sequences of 19 *Sb* and 60 *SB* males from two distinct South American *S. invicta* populations indicates that a supergene identical in structure and content to the one found in the invasive range in the United States also mediates polygyny in

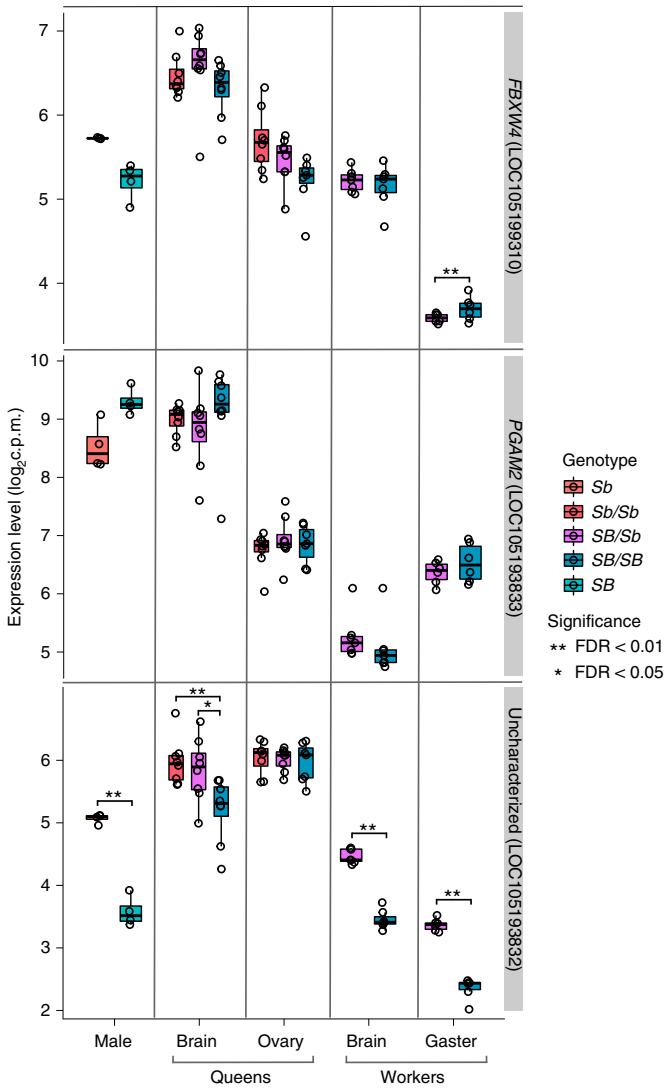



**Fig. 2 | Expression of three protein-coding genes interrupted by supergene inversion breakpoints in *S. invicta*.** Gene expression levels from RNA-seq data are plotted in c.p.m. mapped reads for samples that differ in their social chromosome genotype, including whole bodies of *Sb* and *SB* males ($n = 4$/haplotype), brains and ovaries of *Sb/Sb*, *SB/Sb* and *SB/SB* pre-reproductive queens ($n = 7$–8/genotype) and brains and gasters (abdomens) of *SB/Sb* and *SB/SB* workers ($n = 6$–8/genotype). The box ranges from the first (Q1) to the third quartile (Q3) of the distribution and represents the interquartile range (IQR). A line across the box indicates the median. The whiskers are lines extending from Q1 and Q3 to end points that are defined as the most extreme data points within Q1 − 1.5 × IQR and Q3 + 1.5 × IQR, respectively. Each expression value is represented by a hollow circle. Significance designations refer to *P* values following FDR correction as described in the Methods.

native populations of this species. Specifically, all native *Sb* males harbour the same three inversions characterizing the *Sb* haplotype in the United States, judging from anomalous read pair (ARP) analyses (Extended Data Fig. 4). Moreover, all of these males were collected from confirmed polygyne colonies (19 colonies from Brazil, Argentina and Uruguay; Supplementary Table 5), consistent with previous work showing a perfect association between polygyny and the presence of allele *Gp-9*$^{b}$ (which is diagnostic for *Sb*) in native *S. invicta* colonies[19,20]. In contrast, monogyne colonies ($n = 34$) produced only *SB* males and none of these males (nor any of the 26 *SB* males from polygyne or uncharacterized colonies) harboured any of the three inversions (Extended Data Fig. 4). Thus, the *Sb* haplotype always carries the same three inversions, invariably bears allele *Gp-9*$^{b}$ and is responsible for regulation of colony social organization in native as well as invasive populations of *S. invicta*.

Recombination suppression is crucial in the evolution of supergenes because it stably preserves complementary variants at multiple genes[8,21]. We therefore calculated the extent of LD across the supergene and over the rest of the genome for pooled *Sb* and *SB* males of native *S. invicta* to examine the effects of recombination suppression attributable to the supergene inversions. LD along the 11.4-Mb segment on the distal side of the chromosome 16 centromere was greatly elevated compared to the rest of the genome (mean $r^2 = 0.35$ versus 0.03; Mann–Whitney test, $P < 0.001$; Extended Data Fig. 5a), as was LD along the centromere itself (mean $r^2 = 0.11$). LD along the larger interval, which corresponds closely to the position of the three inversions, also is markedly higher among *Sb* than among *SB* males (mean $r^2 = 0.42$ versus 0.02; Mann–Whitney test, $P < 0.001$; Extended Data Fig. 5b,c), suggesting suppression of effective recombination in *Sb* homozygotes as well as in *SB/Sb* heterozygotes, in contrast to the free recombination assumed to occur among *SB* homozygotes. Such restriction of effective recombination between *Sb* haplotypes probably stems from a combination of the recessive lethality of some of these haplotypes[22,23] as well as structural constraints or other factors limiting crossing-over in inversion homokaryotypes[24].

Because reduced recombination can promote sequence differentiation[25], we quantified both the frequency of fixed differences ($d_f$) and magnitude of nucleotide divergence ($d_{XY}$) between *Sb* and *SB* males. Both of these values were much higher for the supergene region than for the rest of the social chromosome ($d_f$: Fig. 3a; $d_{XY}$: Extended Data Fig. 6a; Mann–Whitney U-tests, both $P < 0.01$). The abrupt increase in $d_f$ and $d_{XY}$ values near position 12.5 Mb and decrease near 24.0 Mb on chromosome 16 match closely the locations of the distal breakpoint of *In(16)1* and proximal breakpoint of *In(16)3*, suggesting that the suppression of recombination (manifested as high LD) and consequent high sequence divergence between the *Sb* and *SB* haplotypes are largely attributable to the three inversions.

Social polymorphism has been shown to be associated with variation at the *Gp-9* locus in several other fire ant species[19,20] and was suggested to be regulated by a supergene in the fire ants *S. richteri* and *S. quinquecuspis*[26]. We sequenced multiple males from each of the five South American species which, together with *S. invicta* and *S. quinquecuspis*, comprise the so-called 'socially polymorphic clade'[19]—*S. macdonaghi* ($n = 3$), *S. megergates* ($n = 7$), *S. richteri* ($n = 56$) and the undescribed *S. AdRX* ($n = 16$) and *S. nr. interrupta* ($n = 4$). In each of these five species, we again observed a high level of LD (Extended Data Fig. 7) and elevated sequence divergence (as measured by both $d_f$ and $d_{XY}$; Fig. 3 and Extended Data Fig. 6) between conspecific *Gp-9*$^{b}$ and *Gp-9*$^{B}$ males in the region of chromosome 16 corresponding to the *S. invicta* supergene. (We note that the small sample sizes for *Sb* haplotypes in *S. macdonaghi*, *S. megergates* and *S. nr. interrupta* lead to somewhat inflated values of $d_f$ both within and outside of the supergene because some intra-haplotype polymorphisms are not identified as such, owing to sampling error.) Importantly, no *Gp-9*$^{b}$ (*Sb*) males of any of the six species had ARP reads connecting downstream and upstream regions adjacent to the breakpoints of the three inversions on the *Sb* reference genome (Extended Data Fig. 4a–c) and, similarly, no *Gp-9*$^{B}$ (*SB*) males had ARP reads connecting analogous regions on the *SB* reference (Extended Data Fig. 4d–f). Together, these data demonstrate that all of the socially polymorphic fire ant species we studied share a homologous inversion-based supergene that is associated with strong suppression of recombination, pronounced sequence divergence between the alternate haplotypes and polygyne

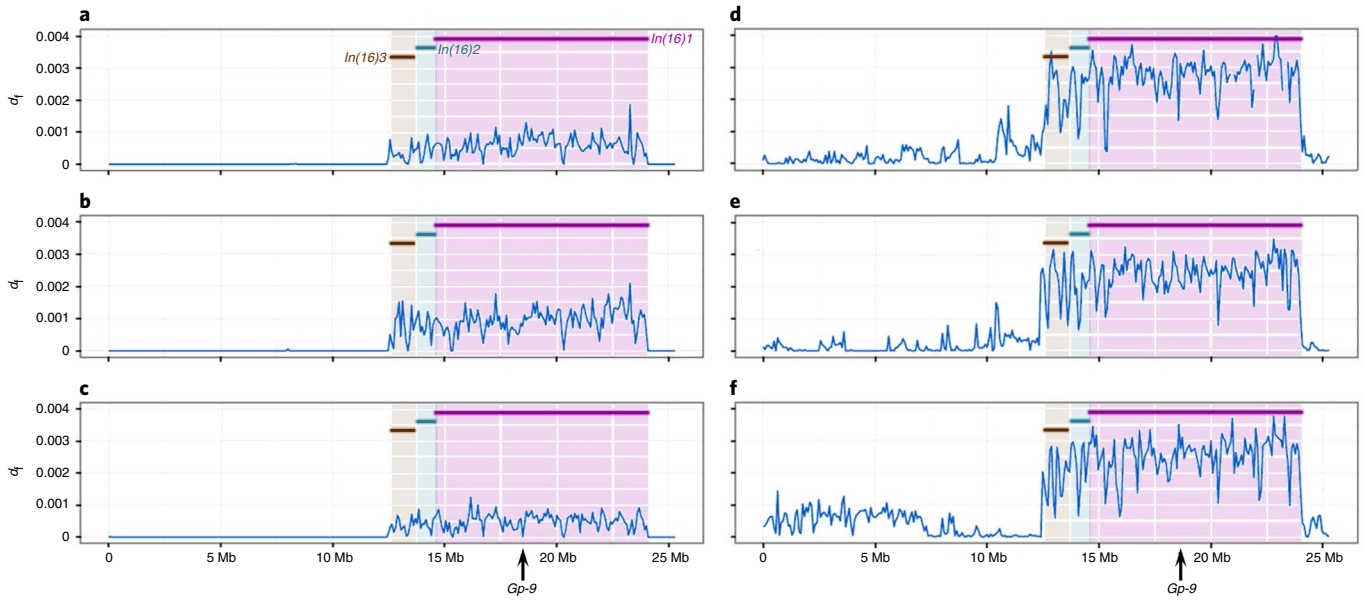



**Fig. 3 | Fixed difference ($d_f$) values across the social chromosome between conspecific *Sb* and *SB* males of six socially polymorphic fire ant species estimated using 50-kb non-overlapping sliding windows. a–f**, Fixed differences for the six species: *S. invicta* (**a**), *S. richteri* (**b**), *S. AdRX* (**c**), *S. nr. interrupta* (**d**), *S. megergates* (**e**) and *S. macdonaghi* (**f**). The boundaries of the differently coloured intervals correspond to the breakpoints of the three inversions, with the inversions depicted by thick horizontal coloured lines. The *x* axes represent physical position along chr16, with the location of gene *Gp-9* indicated. In each of the six species, $d_f$ was significantly higher between the *Sb* and *SB* haplotypes than between the sequences of *Sb* and *SB* males across the rest of the social chromosome (Mann–Whitney U-tests, all $P < 0.001$).

social organization. This confirms previous hypotheses that variation in colony social structure has a common genetic basis in the socially polymorphic clade of South American fire ants[4].

The evolutionary origin of haplotype *Sb* was explored further by conducting a phylogenomic analysis of the socially polymorphic species using SNPs located across all of the genome except the supergene region ($n = 183$ males, Supplementary Table 5). The resulting well-resolved and highly supported phylogeny (Fig. 4a) features four major lineages: (1) the geographically widespread (native range) and highly invasive *S. invicta*, (2) its undescribed sister species, *S. AdRX*[27], (3) the relatively widespread and moderately invasive *S. richteri* and (4) a cluster of three fairly narrow endemics (*S. nr. interrupta*, *S. megergates* and *S. macdonaghi*). Phylogenetic trees inferred using only supergene SNPs differed strikingly from the species tree, with the *Sb* haplotypes from all six species invariably forming a strongly supported clade that appears to have originated either just before the radiation producing the known socially polymorphic species (Fig. 4b) (estimated at ~0.51 million years ago; see Extended Data Fig. 8c) or shortly thereafter (Extended Data Fig. 8a,b). Monophyly of the *Sb* haplotype group is consistent with the conclusion reached earlier that the *Gp-9*$^b$ allele assemblage forms a uniquely derived monophyletic group within the *Gp-9* gene tree in fire ants[3,19,20].

If the *Sb* haplotype lineage originated after the first speciation event, then its presence in all of the species in the socially polymorphic clade would require subsequent hybridization and introgression of *Sb* across species or their stem lineages. As expected, the topology of the *SB* haplotypes is highly congruent with the species relationships; on the other hand, the topology of the *Sb* haplotypes bears less resemblance to the species tree, possibly due to such introgression, selection acting on *Sb* haplotypes[28], confinement of *Sb* transmission to only one social form or occasional intraspecific recombination (gene flux) between *Sb* and *SB* haplotypes.

The potential occurrence of gene flux is supported by the recent finding of rare *SB*/*Sb* recombinants in embryo progenies of *S. invicta*[15]. Therefore, we undertook formal analyses of the extent to which the three supergene inversions have acted historically as barriers to recombination (that is, to gene flux between supergene haplotypes and homologous regions of the wild-type conspecific chromosome). ABBA/BABA tests (*D*-statistics) revealed that mutations shared between conspecific *SB* and *Sb* haplotypes significantly exceed the number expected from only recurrent convergent mutation in all six studied species (Fig. 5). Moreover, data from the three best-sampled species suggest that gene flux has occurred in both directions (*SB* → *Sb*, *Sb* → *SB*). Direct comparison of shared polymorphic sites (SPS) between the *SB* and *Sb* haplotypes provides additional evidence for historical gene flux, as follows. Given its unique origin, the *Sb* haplotype initially must have been monomorphic. In the absence of gene flux, SPS between *Sb* and *SB* haplotypes can only have arisen through convergent mutations and, therefore, should be rare. Contrary to this expectation, >10% of the polymorphic sites in *Sb* haplotypes also were polymorphic in the *SB* haplotypes, a figure far too high to be attributed solely to recurrent mutation but compatible with occasional historical gene flux between the *SB* and *Sb* haplotypes.

Interhaplotype gene flux is predicted to lead to a negative association between the proportion of SPS and interspecific divergence time when comparing *Sb* and *SB* haplotypes of different species, because more recently diverged species have a longer history of shared ancestry since the origin of *Sb* and, thus, extended opportunities for *Sb*-*SB* gene flux compared to more distantly related pairs of species that diverged earlier. Consistent with this prediction, SPS were significantly more common within than between each of the two main socially polymorphic clades (*S. invicta*-*richteri*-*AdRX* and *S. macdonaghi*-*megergates*-*nr. interrupta*; Fig. 6a; Mann–Whitney U-tests, both $P < 0.001$) and there was a strong negative relationship between the proportion of SPS and interspecific nucleotide divergence ($d_{XY}$) for the three best-sampled species (Fig. 6b, Mantel test, $P < 0.001$). These findings, together with the earlier progeny-study results[15], indicate that recombination between conspecific *Sb* and *SB* haplotypes is not entirely suppressed by the three inversions on *Sb* but has occurred at a low rate following diversification of the socially polymorphic clade.

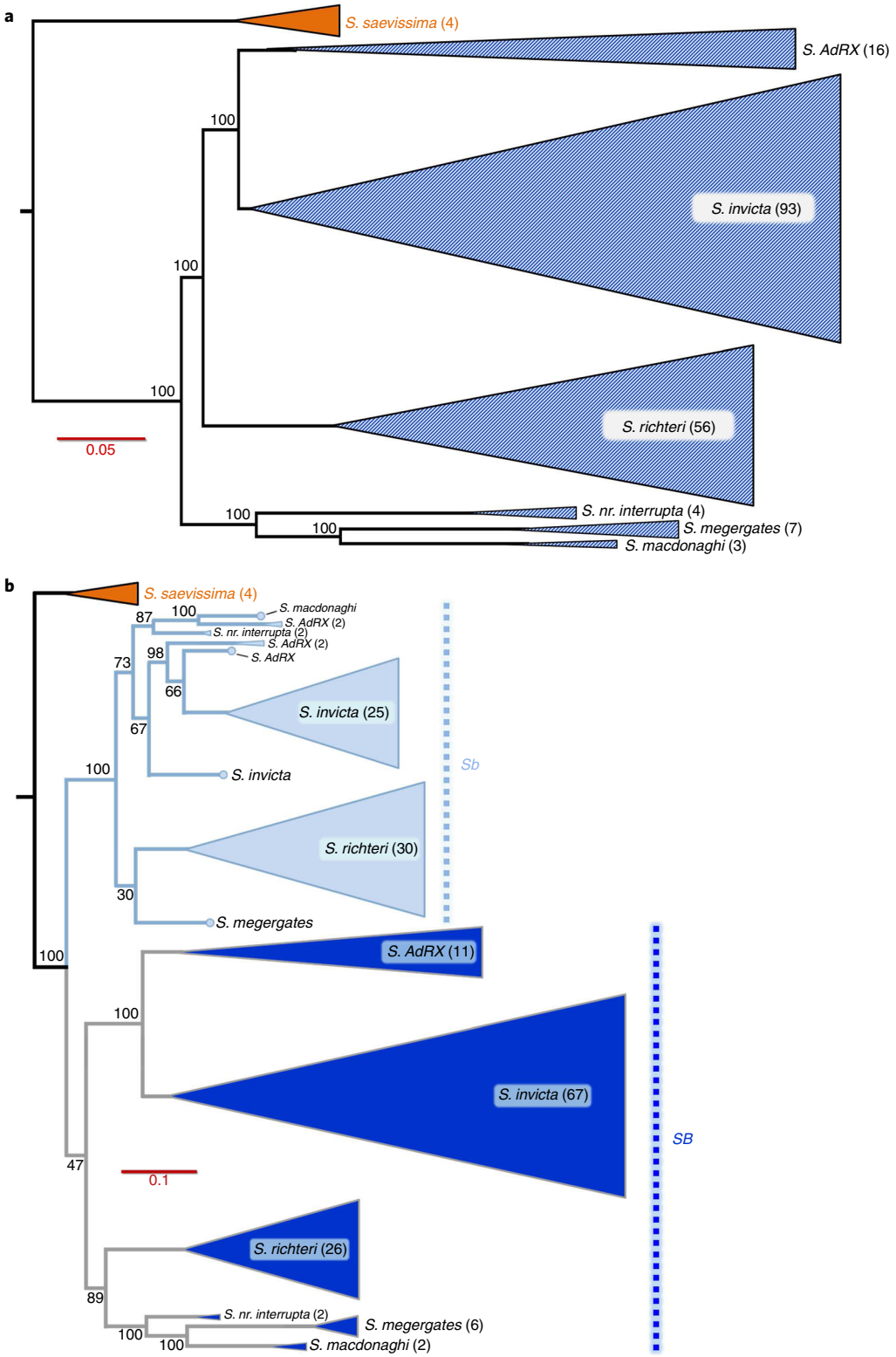



**Fig. 4 | Phylogenies of South American fire ant species in the socially polymorphic clade and of the supergene region in these species.** ML phylogenetic trees feature triangular terminals that represent collapsed nodes of related sequences (100% bootstrap support); height is proportional to the number of sequenced males (shown in parentheses after species name) comprising each collapsed node. Bootstrap support values are shown at relevant nodes. Trees were rooted using sequences from the outgroup species *S. saevissima*, which lacks the chr16 inversions. The red scale bars are substitutions per site. **a**, Species tree of the six species based on 12,237,341 genome-wide SNPs, excluding those in the supergene. Cross-hatching indicates the presence of both monogyny and polygyny and of both the *Sb* and *SB* haplotypes in a species. **b**, Tree of the *Sb* and *SB* haplotypes of chr16 in the six species based on 610,247 SNPs and that accounts for LD, using a pruning threshold of 0.8. The dark and light blue colours represent sequences recovered from *SB* and *Sb* males, respectively. The vertical dotted lines highlight the positions of two major haplotype groups in the phylogeny (the *SB* and *Sb* clades).

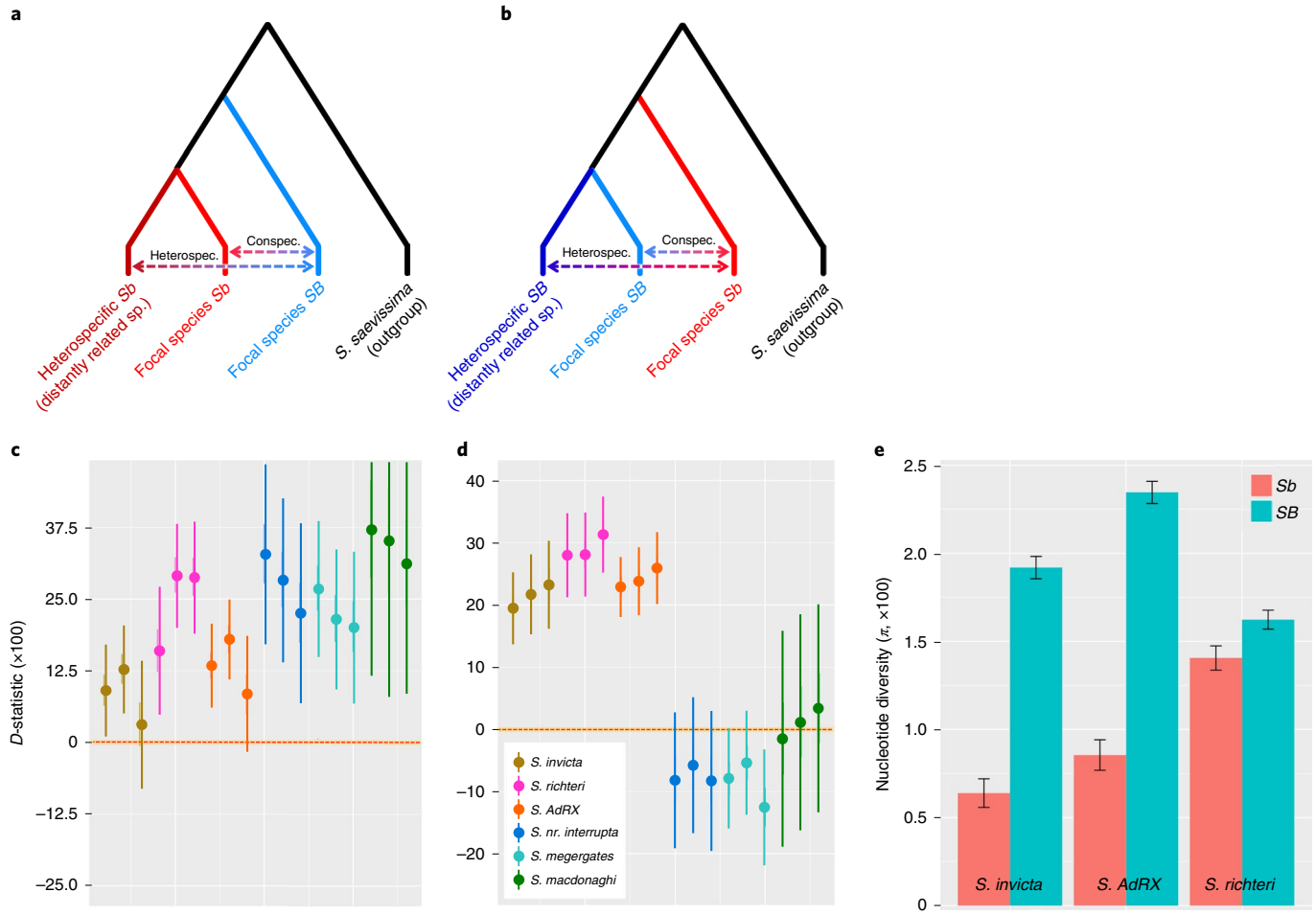



**Fig. 5 | Results of ABBA/BABA tests (*D*-statistics) for recombination (gene flux) between conspecific *Sb* and *SB* haplotypes and estimates of nucleotide diversity ($\pi$). a**,**b**, ABBA/BABA tests were performed for each of the six socially polymorphic *Solenopsis* study species to detect gene flux from *SB* into conspecific *Sb* haplotypes by comparing focal *SB* males to conspecific and heterospecific *Sb* males (**a**) and to detect gene flux in the opposite direction by comparing focal *Sb* males to conspecific and heterospecific *SB* males (**b**). **c**, Gene flux from *SB* into conspecific *Sb* haplotypes; for each of the six species, *D*-statistics from the ABBA/BABA tests were computed by using *Sb* haplotypes from each of the three most distantly related species for the heterospecific comparisons (that is, for each species there are three points of the same colour indicating the values of these three estimates). Error bars indicate ±3 s.e.m. estimated with a weighted-block jackknife approach using 50-kb blocks of the complete supergene region; non-overlap with zero signifies a statistically significant *D*-value (evidence for gene flux). **d**, Gene flux from *Sb* into conspecific *SB* haplotypes was measured as in **c** but with *SB* males used as referents to estimate gene flow in the opposite direction. **e**, Nucleotide diversity ($\pi$) within the supergene region for *Sb* and *SB* males of the three best-sampled species (native ranges); error bars indicate the 95% CIs about the means derived from resampling of 228 non-overlapping 50-kb windows. Nucleotide diversity was significantly lower in *Sb* than *SB* males in each of the species (*t*-tests, all $P < 0.001$). The relatively modest difference in diversity between *Sb* and *SB* in *S. invicta* (as well as the other two species), stands in stark contrast to the extreme reduction in *Sb* diversity relative to *SB* diversity in invasive *S. invicta* portrayed in Pracana et al.[28].

The moderate levels of *Sb* diversity we document (one-third the level of *SB* diversity in *S. invicta*; Fig. 5e) presumably reflect in some measure the effects of such low historical levels of interhaplotype gene flux. The very different finding by Pracana et al.[28] that *Sb* diversity is only a minute fraction (<1%) of *SB* diversity can be explained by the fact that only a few samples from a recently bottlenecked invasive *S. invicta* population were analysed in that study. The presence of such dramatic differences in *Sb* diversity between native and introduced populations predicts corresponding differences in polygyne trait variation and the responsiveness of such traits to selection, an idea supported by earlier findings of strong differences in major features of social organization between Argentine and US polygyne populations that mirror differences in overall genetic diversity between the two ranges[29,30].

Large supergenes and sex chromosomes that comprise several inversions are often characterized by 'evolutionary strata', segments of varying sequence divergence that reflect differences in evolutionary age of the inversions[31,32]. Previous work based on a limited number of *S. invicta* samples from the invasive range showed no evidence for strata of differentiation along the supergene, leading the authors to propose that a single event may have led to suppression of recombination over the entire supergene[28]. To test this proposal, we examined between-haplotype sequence differentiation at each of the three inversion regions for the six socially polymorphic study species as well as for pooled data for subsets of species (clades) using two different metrics. The rationale for these analyses is the expectation that longer periods since inversion of a segment are reflected in increased divergence between it and conspecific wild-type homologues (owing to the reduction in recombination). Distributions of values for both statistics $d_{XY}$ and $d_S$ among the inversions generally are as predicted by the hypothesized order of emergence *In(16)1→In(16)2→In(16)3* for

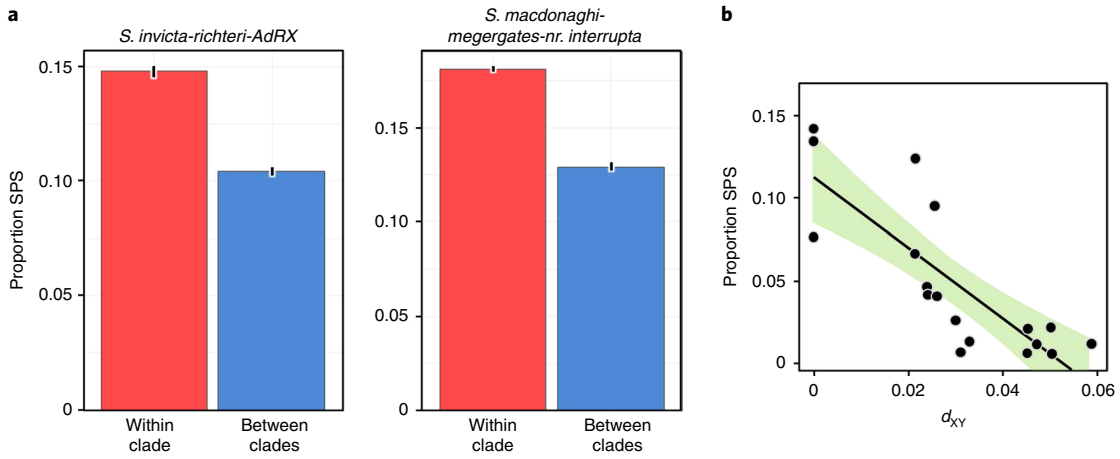



**Fig. 6 | SPS between *Sb* and *SB* haplotypes of different species. a**, Left panel shows SPS between *Sb* haplotypes from the *S. invicta-richteri-AdRX* clade and *SB* haplotypes from the same clade (red bar) or from the *S. macdonaghi-megergates-nr. interrupta* clade (blue bar); right panel shows SPS between *Sb* haplotypes from the *S. macdonaghi-megergates-nr. interrupta* clade and *SB* haplotypes from the same clade (red bar) or from the *S. invicta-richteri-AdRX* clade (blue bar). Error bars represent 95% CIs from 1,000 resampling iterations. **b**, SPS between *Sb* haplotypes from the three best-sampled species (*S. invicta, S. richteri* and *S. AdRX*) and *SB* haplotypes from all six socially polymorphic species studied plotted against interspecific nucleotide divergence ($d_{XY}$). The green zone defines the 95% CI for predictions from a negative linear regression model ($r = -0.82$). Historical gene flux is predicted to lead to a negative relationship between the proportion of SPS and divergence time when comparing *Sb* and *SB* haplotypes of different species because more closely related (recently diverged) species have a longer history of shared ancestry since *Sb* originated and, thus, greater opportunities for gene flux between *SB* and *Sb*, than more distantly related species. This prediction is supported by the greater within-clade than between-clades SPS as well as the negative correlation between proportion of SPS and $d_{XY}$.

the individual species, a pattern recapitulated in comparisons using pooled data (Extended Data Fig. 9). The largely congruent single-species patterns are statistically significant in aggregate (probabilities of no between-inversion differences in haplotype divergence: $P < 0.05$ for *In(16)1* versus *In(16)2*, $P < 0.001$ for *In(16)1* versus *In(16)3* and $P < 0.005$ for *In(16)2* versus *In(16)3*; Fisher method of combining one-tail Mann–Whitney U-test probabilities across species and metrics). These findings corroborate Huang et al.'s conclusion[11] that *In(16)1* originated before *In(16)2*, based on the location/orientation of the doubly inverted fragment and further indicate that the centromere-bridging third inversion is the most recent to have appeared.

In summary, our study shows that variation in social organization in the six focal fire ant species we studied is controlled by a supergene that makes up half of the social chromosome. The *Sb* supergene haplotype evidently evolved via sequential incorporation of three contiguous inversions. The first inversion to appear caused minor truncation of the coding sequence of one protein-coding gene, while the second inversion resulted in broadly manifested increases in expression levels at a second such gene. The third, most recently acquired, inversion expanded the supergene boundaries by bridging the centromere and the older inversions. These structural novelties induced restricted recombination between *Sb* and *SB* haplotypes, which presumably preserved in alternate sequestered haplotypes the complementary genetic variants at many coadapted genes that influence various differences in dispersal and reproductive strategies between the monogyne and polygyne forms. Important issues remaining to be resolved are the manner in which such complementary alleles initially were assembled in the derived *Sb* haplotype[33] and the identities of supergene loci that drove selection to maintain a stable polymorphism in the face of potential Hill–Robertson interference across the linked genes[31,34,35]. Because the *Sb* haplotype lineage originated near the time of radiation of the socially polymorphic species, it evidently has been maintained through multiple speciation events, leading to a widespread trans-species polymorphism with the *SB* haplotypes. Such polymorphisms, which are rare in nature, reflect the action of persistent balancing selection[36,37] which, in the case of fire ants, probably is related to the ecological advantages of each social form in different circumstances[38] and to the *SB* haplotype opposing the selfish genetic tendencies of *Sb*[15]. Finally, our analyses indicate that gene flux (recombination) is not entirely suppressed between the *Sb* and *SB* haplotypes, a result with important implications. Specifically, although crossing-over is expected to be limited in inversion heterozygotes[39], gene conversion events may actually be accelerated and yield modest between-haplotype gene flux[40]. Other studies in diverse taxa also have documented sporadic gene flux occurring by double crossovers or gene conversion events between large inverted and wild-type genomic regions analogous to supergene systems[41], even between the X and Y chromosomes in some vertebrates[42]. We therefore suggest that low levels of recombination and/or gene conversion may play an underappreciated role in preventing rapid degeneration of supergenes, by allowing novel variants to infiltrate inversions at a rate insufficient to cause significant decay of LD but sufficient to forestall the effects of Muller's ratchet[43], thereby contributing to the persistence of a genetic architecture underlying many complex adaptive polymorphisms.

## Methods

**Sample collection.** Samples were collected from the native ranges of seven fire ant species in South America during collection trips from 1990 to 2015. These were *S. invicta*, *S. macdonaghi*, *S. megergates* and *S. richteri*, as well as the undescribed *S. AdRX* and *S. nr. interrupta*, all members of the socially polymorphic clade of South American fire ants; *S. saevissima* was sampled as an outgroup species. A total of 169 specimens from these samples were newly sequenced (Supplementary Table 5). We used haploid males exclusively for genome sequencing to directly infer haplotypes. We used only a single male from each monogyne colony to avoid sequencing genetically related individuals and, whenever possible, we sequenced one *Gp-9*$^{b}$ and one *Gp-9*$^{B}$ male from each polygyne colony. In addition, we used the data from 14 previously sequenced *S. invicta* males from the invasive (United States) range[5].

**Determination of social form.** Twelve workers from each colony were genotyped at ten highly variable microsatellite loci to verify colony social form by means of identifying a single or multiple matrilines[19]. Additionally, each male specimen was typed with a TaqMan allelic discrimination assay at three key codon positions considered to be diagnostic for *b*-like alleles of *Gp-9* (refs. [44,45]), with the objective to sample as much of the diversity of *Gp-9* alleles and *Sb* haplotypes as possible.

All males with one or more of the *Gp-9*$^{b}$ diagnostic codons and possessing the three inversions were classified as *Gp-9*$^{b}$ (*Sb*) individuals, while all those lacking the inversions were classified as *Gp-9*$^{B}$ (*SB*) individuals.

**Illumina HiSeq sequencing.** Standard Illumina protocols (TruSeq DNA) were used to prepare the paired-end libraries. Briefly, genomic DNA was isolated from the whole bodies of single male ants. After fragmentation using a Covaris instrument, the short insert size DNA fragments (target size: 500–550 base pairs, bp) were end repaired and ligated to the Illumina Pair-End Sequencing adaptors. Ligated products were polymerase chain reaction (PCR)-amplified (15 cycles) and each genome library was then sequenced to at least 10× sequencing depth (Supplementary Table 5).

**Reference genome assembly using PacBio 10-kb long reads.** Sequencing data were derived from a single *S. invicta* male pupa with the *SB* haplotype from a polygyne colony from Florida, United States, and this served as the basis for the *SB* reference genome assembly (National Center for Biotechnology Information, NCBI, genome ID *Solenopsis_invicta_ M01_SB*). DNA was isolated using the Genomic-tip 20/G extraction Kit (Qiagen) following the manufacturer's instructions. Genomic DNA was sheared to a size range of 15–50 kb using G-tubes (Covaris) and enzymatically repaired and converted into SMRTbell template libraries as recommended by Pacific Biosciences. In brief, hairpin adaptors were ligated, after which the remaining damaged DNA fragments and those without adaptors at both ends were eliminated by digestion with exonucleases. The resulting SMRTcell templates were size-selected to 15–50 kb by Blue Pippin electrophoresis (Sage Sciences) and sequenced on a PacBio RS II instrument using P6-C4 sequencing chemistry. Data from 40 SMRT cells were collected and 4,061,662 reads (43,553,361,954 bases) were obtained in total. The CANU assembly pipeline v.1.4 (ref. [46]) was used to perform correction of the reads, trim and assemble. The raw assembly includes 1,447 contigs (N50 size of 956,625 bp). We used JCVI (v.0.7.1, ref. [47]) to anchor and orient these contigs using three equally weighted linkage maps constructed from single-queen (*Gp-9*$^{BB}$) families[5] with restriction site-associated DNA sequencing (RADseq) SNPs. A total of 64.5% of the contigs (264,123,748 bp) were anchored into 16 linkage groups (chromosomes), with 36.5% of the contigs (144,930,346 bp) remaining unplaced. The total length of contigs located on the social chromosome was 25,360,913 bp, which is 31.4% more than the former gnG assembly version (19,291,722 bp)[5].

To identify the inversions that distinguish *Sb* from *SB* haplotypes, we sequenced a single *S. invicta Sb* male pupa from the same colony as above to generate the *Sb* reference genome (NCBI genome ID *Solenopsis_invicta_M02_Sb*), using the identical PacBio RS II protocol. Data from 40 SMRT cells were collected and 3,684,238 reads (46,474,046,294 bp) were obtained in total. The CANU assembly procedure generated 1,372 contigs (N50 size of 1,228,881 bp). We anchored and oriented these contigs using four equally weighted linkage maps from families headed by single polygyne (*Gp-9*$^{bb}$) queens that were previously generated[5] using RADseq. A total of 86.8% of the contigs (337,734,421 bp) were anchored into 16 linkage groups, with 13.2% of the contigs (51,415,094 bp) remaining unplaced.

**Inversions in the supergene.** To identify the inversions defining the supergene, a pairwise alignment between the *Sb* and *SB* reference genomes was performed using LAST V531 (ref. [48]). The last-dotplot script was used to generate a dot plot of the pairwise alignment. To fully characterize the inversion breakpoints, we manually checked the pairwise alignment. The closest nucleotide position separating the alignment into two distant regions along the *SB* haplotype was defined as the breakpoint position. We included 200-bp segments upstream and downstream of the breakpoints to calculate sequence similarity between the *Sb* and *SB* haplotypes around the breakpoints.

To test whether the inversions disrupted any protein-coding gene, we downloaded the *S. invicta* Annotation Release 100 from NCBI (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Solenopsis_invicta/100/) and mapped 14,453 protein-coding genes to the *SB* reference genome. The genome coordinates of these protein-coding genes were compared to the inversion breakpoint coordinates to identify candidate interrupted genes.

To investigate the potential functions of *PGAM2* and *FBXW4*, we compared each protein sequence with its putative orthologues, identified by reciprocal best BLAST among available protein sequences for *Drosophila melanogaster, Danio rerio, Mus musculus* and *Homo sapiens* downloaded from NCBI. The catalytic core and typical domains were predicted for *PGAM2* and *FBXW4*, respectively, by InterPro and NCBI. Consensus sequences were produced using WebLogo[49].

**Analysis of expression of genes interrupted by inversion breakpoints.** Males and workers were sampled from *S. invicta* colonies collected in Florida, United States, and reared for 1 month under standard laboratory conditions before any experimental manipulations. Males were sampled within 1 h after emergence from the pupal stage[50]. Workers, on reaching the pupal stage, were transferred to polygyne host colonies and maintained for 14 d before sampling. Pre-reproductive queens were sampled as they were departing on their mating flights at multiple sites in Georgia, United States. Males and workers were typed with a TaqMan allelic discrimination assay[45] and queens were typed with a modified multiplex PCR assay[51] to determine *Gp-9* haplotypes (males) or genotypes. RNA was extracted from whole bodies of males and from worker brains and gasters (abdomens) using a Trizol protocol. RNA was extracted from queen brains and ovaries using the RNeasy Micro Kit and RNeasy Mini Kit, respectively. Sequencing libraries were prepared from male bodies and worker brains using the SMARTer v.3 kit for polyA-selected mRNA, from worker gasters using a KAPA stranded RNA-sequencing kit and from queen brains and ovaries using the SMART-seq2 protocol[52]. All libraries were sequenced using standard Illumina protocols. Sequencing reads are available from the NCBI (PRJNA421367).

RNA-sequencing reads were aligned to the *Solenopsis_invicta_M01_SB* reference genome using STAR's 2-pass approach with default parameters[53]. Gene counts were then generated on the basis of the Refseq gene models mapped to the *Solenopsis_invicta_M01_SB* reference assembly. Differential expression and gene counts per million (c.p.m. reads) values were computed using edgeR[54]. Additionally, we leveraged fixed, inversion-defining SNPs generated using DNA sequences of *S. invicta* males from the native range to compute allele-specific expression at the three inversion breakpoint genes. In heterozygous workers and queens, allele-specific read counts were generated using GATK's ASEReadCounter[55,56] and allele-specific expression differences were computed using edgeR. Finally, we used DEX-seq to quantify differential exon usage between individuals with and without the supergene in each of our sample types. In all analyses, differences were considered significant with an FDR-corrected $P < 0.05$ (ref. [57]).

**Inversions in the other socially polymorphic South American fire ant species.** Raw paired-end Illumina sequence reads (PE reads) from 169 males of the six socially polymorphic species were mapped to both the *Sb* and *SB* reference genomes. The raw PE reads that aligned inappropriately to the 800-bp windows around both the proximal and distal inversion breakpoints on the *Sb* or *SB* reference genomes were defined as ARPs. The number of ARPs within each sample was counted separately for inversions *In(16)1, In(16)2* and *In(16)3*. One sample of *S. AdRX* and three samples of *S. richteri*, *S. nr. interrupta* and *S. macdonaghi* lacked ARPs directly anchoring the two breakpoints of *In(16)1* and *In(16)3*. Thus, targeted local assembly of breakpoint sequences (TIGRA)[58] software was used to perform local assembly. The bam files containing raw reads from the these samples were extracted as input for TIGRA and the genome coordinates for the *In(16)3* proximal and distal breakpoints were marked as regions of interest for local assembly. The sets of reads mapped or partially mapped to these regions were assembled into contigs. We further mapped these assembled contigs to the *Sb* and *SB* genomes to confirm that they connected the *In(16)*1 and *In(16)3* breakpoints respectively.

**Genotype calling and LD analyses.** Illumina paired-end reads from the 190 fire ant male genomes were filtered (minimum quality score 10) and aligned to the reference *SB* genome using bwa v.0.7.10 (ref. [59]). The mapping results were sorted, duplicate marked and indexed using SAMtools v.0.1.19 (ref. [60]). Because all assembled genomes derived from haploid male specimens, we used the haplotype-based variant detector Freebayes v.0.9.21 (ref. [61]) to call haplotypes across all individuals with the following parameters: ploidy 1, min-mapping-quality 1, mismatch-base-quality-threshold 10. We detected 2,337,243 variants (SNPs) along the social chromosome and 73,406,103 along the remaining 15 chromosomes.

For the linkage disequilibrium (LD; gametic disequilibrium) analyses, we first used VCFtools v.0.1.14 9 (ref. [62]) to filter loci (missing data < 10%, minor allele frequency > 0.2, sites quality value > 1,000, maximum alleles = 2) in each species. We next used this software to compute $r^2$ statistics for pairwise LD. The R package gg.ldplot (https://github.com/timknut/gg.ldplot) was used to make dot plots to visualize pairwise LD (Extended Data Fig. 5 and Extended Data Fig. 7).

**Sequence divergence between *Sb* and *SB* haplotypes.** To analyse patterns of differentiation between the *Sb* and *SB* haplotypes in each species, we used VCFtools[62] to calculate SNP frequencies then used the script popgenWindows.py (https://github.com/simonhmartin/genomics_general) to compute nucleotide divergence ($d_{XY}$) using non-overlapping 50-kb sliding windows along chromosome 16. We next used VCFtools[62] to calculate SNP frequencies and manually computed fixed differences ($d_f$) between *Sb* and *SB* haplotypes in each species using non-overlapping 50-kb sliding windows along chromosome 16. Finally, we used the script popgenWindows.py to compute nucleotide diversity statistics ($\pi$) using non-overlapping 50-kb sliding windows along chromosome 16 for *Sb* and *SB* males of the three best-sampled species.

**Phylogenomic analyses.** We constructed separate phylogenetic trees for the supergene region on chr16 and for the remainder of the genome using all six study species, with *S. saevissima* included as the outgroup in all analyses. This latter species falls outside of the socially polymorphic South American fire ant clade but is placed along with that clade in the *S. saevissima* species-group[63]; *S. saevissima* lacks the three chr16 inversions characterizing the *Sb* haplotype, thus confirming the *SB*-like chr16 architecture as the ancestral condition for the socially polymorphic clade. For the whole-genome phylogenomic analyses, we constructed a maximum-likelihood (ML) tree based on concatenation of 12,237,341 SNPs retained after filtering (supergene region excluded, bi-allelic sites, quality > 30). SNPhylo 3.69 (ref. [64]) was used to construct the whole-genome tree

with the following parameter settings: LD threshold 0.1, missing rate 0.1, number of bootstrap replicates 1,000. For the supergene analysis, we used RAxML v.8.1.2 9 (ref. [65]) to construct an ML tree for the 610,247 SNPs with the GTRGAMMA model and 1,000 bootstrap replicates. We then used SNPhylo v.3.69 (ref. [64]) to construct two additional ML trees that corrected for the strong LD between (non-independence of) supergene SNPs[66]; for these analyses, the LD threshold was set at either 0.5 or 0.8, the missing rate at 0.1 and the number of bootstrap replicates at 1,000. The best supported supergene phylogeny, with respect to the relationship between *Sb* and *SB* haplotype lineages, was derived from the analysis with the highest level of accommodation of supergene LD (LD threshold = 0.8). This tree features 100% bootstrap support between the *Sb* clade (recovered with 100% support in all three trees) and its sister group, in this case a monophyletic *SB* group. The two alternative trees featured 92% or 84% bootstrap support for the relationship between the *Sb* clade and its hypothesized sister group (in each case, a subset of the *SB* haplotypes).

**Recombination (gene flux) between the *Sb* and *SB* haplotypes.** The VCF file containing 610,247 SNPs within the supergene region was converted to the EIGENSTRAT format. ADMIXTOOLS v.4.1 (ref. [67]) was used to conduct ABBA/BABA tests[68] to infer gene flux between conspecific *Sb* and *SB* haplotypes. We used *S. saevissima* as the outgroup and examined the numbers of derived alleles shared between three ingroup populations by calculating *D*-statistics. Significance of the *D*-statistics was assessed with a block jackknife procedure using a *z* score of three standard errors as the threshold[67].

As a further means of assessing gene flux, we analysed patterns of SPS between the *Sb* and *SB* haplotypes. We first used VCFtools to calculate frequencies of each variant of the *Sb* and *SB* haplotypes in the focal species or clades. A polymorphic site in both the *Sb* and *SB* haplotypes was classified as an SPS if both haplotypes shared at least two identical bases. The proportion of SPS was defined as the total number of SPS divided by the total number of polymorphic sites within each haplotype. Next, we estimated nucleotide divergence ($d_{XY}$) between paired socially polymorphic species with all genome-wide SNPs (excluding the supergene region on chr16) using the script popgenWindows.py (https://github.com/simonhmartin/genomics_general). For each of the three species for which we generated five or more *Sb* haplotypes (*S. invicta, S. AdRX* and *S. richteri*), we determined the proportion of polymorphic sites in the *Sb* haplotypes that were also polymorphic in the *SB* haplotypes of each of the six species in the socially polymorphic clade (including the focal species). The relationship of the resulting 18 SPS values to the nucleotide divergence between paired species (zero in the case of each focal species) was then determined.

**Times of species divergence and origin of the *Sb* haplotype.** We first estimated the time of origin of the socially polymorphic clade of fire ants and divergence times of the six species using coding sequences of five previously studied nuclear genes (*18S rDNA*, *28S rDNA*, *abdominal-A*, *long-wavelength rhodopsin* and *wingless*);[69] corresponding sequences from *S. xyloni* (a fire ant not included in the *S. saevissima* species-group[63] that was used as the outgroup species) were downloaded from TreeBASE (study accession http://purl.org/phylo/treebase/phylows/study/TB2:S11283; www.treebase.org). Divergence times were estimated using BEAST v.2.4.7 (ref. [70]), with the GTR + G model of substitution and an uncorrelated log-normal relaxed clock model with a Yule process for the prior on tree topologies. For calibration, a log-normal prior distribution was set with an offset equal to the minimum divergence time between *S. xyloni* and *S. invicta* (that is, 4.5 million years;[69] log(mean) = 1.0, log(s.d.) = 1.0). The MCMC chain was run for $50 \times 10^6$ generations and parameters were sampled every 1,000 generations. To accelerate convergence, we used the tree obtained by RAxML as a starting tree and prevented the topology from updating by removing the four operators: Wide-exchange, Narrow-exchange, Wilson–Balding and Subtree-slide. We checked the convergence patterns of the MCMC using Tracer and discarded the first 10% of chain burn-ins before estimating the posterior distributions using TreeAnnotator.

**Order of appearance of the three supergene inversions.** Values of $d_{XY}$ between conspecific *Sb* and *SB* haplotypes were estimated for the three inversion regions for each of the six socially polymorphic species. Means and 95% confidence intervals (CIs) were obtained from 1,000 bootstrap replicates across homologous non-overlapping 50-kb sliding windows ($n = 190$, 16 and 21 windows assigned to inversions *In(16)1*, *In(16)2* and *In(16)3*, respectively, based on the start and end coordinates). The $d_{XY}$ values for each window were averages for all pairs of conspecific *Sb* and *SB* haplotypes. Values of $d_{XY}$ for each inversion region that did not differ significantly between species ($\alpha = 0.05$; Kruskal–Wallis tests followed by Dunn's multiple-comparison tests adjusted using the Benjamini–Hochberg FDR method) were pooled to increase statistical power.

Values of $d_S$ (number of synonymous substitutions per synonymous site) between conspecific *Sb* and *SB* haplotypes were estimated to complement the $d_{XY}$ analyses. SNPs within the coding regions of *Sb* haplotypes were annotated as synonymous or non-synonymous substitutions by the software SnpEff[71] using the reference *SB* haplotype from *S. invicta*, with the total number of synonymous sites calculated using the software PAML[72,73]. The genome assembly of the outgroup species *S. saevissima* (not a member of the socially polymorphic South American fire ant clade) was used to aid in assigning ancestral states and thus in designating each SNP in *Sb* haplotypes as a synonymous or non-synonymous substitution. All coding genes in the *Sb* haplotype with integral open reading frames were assigned to inversion *In(16)1*, *In(16)2* or *In(16)3* ($n = 396$, 33 and 16 genes, respectively). Means and 95% CIs for $d_S$ were derived from 1,000 bootstrap replicates across single homologous genes. Values of $d_S$ for each inversion region that did not differ significantly between species ($\alpha = 0.05$; Kruskal–Wallis tests followed by Dunn's multiple-comparison tests adjusted using the Benjamini–Hochberg FDR method) were pooled to increase statistical power.

Comparisons of $d_{XY}$ and $d_S$ values were used to test for differences among the three inversions in their levels of divergence between the homologous *Sb* and *SB* haplotypes (that is, to assess the presence of evolutionary strata). Order of appearance of the inversions is inferred under the assumption that greater divergence between the haplotypes corresponds to greater age of the inversions.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The genome assembly, gene models and sequence reads are available at the NCBI under the BioProject PRJNA421367.

Received: 10 September 2019; Accepted: 4 December 2019;
Published online: 20 January 2020

## References

1. Ross, K. G. Multilocus evolution in fire ants: effects of selection, gene flow and recombination. *Genetics* **145**, 961–974 (1997).
2. Ross, K. G. & Keller, L. Genetic control of social organization in an ant. *Proc. Natl Acad. Sci. USA* **95**, 14232–14237 (1998).
3. Krieger, M. J. B. & Ross, K. G. Identification of a major gene regulating complex social behavior. *Science* **295**, 328–332 (2002).
4. Gotzek, D. & Ross, K. G. Genetic regulation of colony social organization in fire ants: an integrative overview. *Q. Rev. Biol.* **82**, 201–226 (2007).
5. Wang, J. et al. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493**, 664–668 (2013).
6. Ross, K. & Keller, L. Experimental conversion of colony social organization by manipulation of worker genotype composition in fire ants (*Solenopsis invicta*). *Behav. Ecol. Sociobiol.* **51**, 287–295 (2002).
7. Charlesworth, D. & Charlesworth, B. Theoretical genetics of Batesian mimicry II. Evolution of supergenes. *J. Theor. Biol.* **55**, 305–324 (1975).
8. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Curr. Biol.* **24**, R288–R294 (2014).
9. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity* **113**, 1–8 (2014).
10. Zhang, W., Westerman, E., Nitzany, E., Palmer, S. & Kronforst, M. R. Tracing the origin and evolution of supergene mimicry in butterflies. *Nat. Commun.* **8**, 1269 (2017).
11. Huang, Y.-C., Dang, V. D., Chang, N.-C. & Wang, J. Multiple large inversions and breakpoint rewiring of gene expression in the evolution of the fire ant social supergene. *Proc. R. Soc. B* **285**, 20180221 (2018).
12. Ho, M. S., Tsai, P.-I. & Chien, C.-T. F-box proteins: the key to protein degradation. *J. Biomed. Sci.* **13**, 181–191 (2006).
13. Long, M., Betrán, E., Thornton, K. & Wang, W. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. *Genetics* **129**, 1085–1098 (1991).
14. Huang, Y.-C. et al. Evolution of long centromeres in fire ants. *BMC Evol. Biol.* **16**, 189 (2016).
15. Ross, K. G. & Shoemaker, D. Unexpected patterns of segregation distortion at a selfish supergene in the fire ant *Solenopsis invicta*. *BMC Genet.* **19**, 101 (2018).
16. Jaenike, J. Sex chromosome meiotic drive. *Annu. Rev. Ecol. Syst.* **32**, 25–49 (2001).
17. Fritz, G. N., Vander Meer, R. K. & Preston, C. A. Selective male mortality in the red imported fire ant, *Solenopsis invicta*. *Genetics* **173**, 207–213 (2006).
18. Lawson, L. P., Vander Meer, R. K. & Shoemaker, D. Male reproductive fitness and queen polyandry are linked to variation in the supergene *Gp-9* in the fire ant *Solenopsis invicta*. *Proc. R. Soc. B* **279**, 3217–3222 (2012).
19. Gotzek, D., Shoemaker, D. & Ross, K. G. Molecular variation at a candidate gene implicated in the regulation of fire ant social behavior. *PLoS ONE* **2**, e1088 (2007).
20. Krieger, M. J. B. & Ross, K. G. Molecular evolutionary analyses of the odorant-binding protein gene *Gp-9* in fire ants and other *Solenopsis* species. *Mol. Biol. Evol.* **22**, 2090–2103 (2005).
21. Charlesworth, D. The status of supergenes in the 21st century: recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. *Evol. Appl.* **9**, 74–90 (2016).
22. DeHeer, C. J., Goodisman, M. A. D. & Ross, K. G. Queen dispersal strategies in the multiple-queen form of the fire ant *Solenopsis invicta*. *Am. Nat.* **153**, 660–675 (1999).

23. Hallar, B. L., Krieger, M. J. B. & Ross, K. G. Potential cause of lethality of an allele implicated in social evolution in fire ants. *Genetica* **131**, 69–79 (2007).
24. Remis, M. I. Chromosome polymorphisms in natural populations of the South American grasshopper *Sinipta dalmani*. *Genet. Mol. Biol.* **31**, 42–48 (2008).
25. Campos, J. L., Charlesworth, B. & Haddrill, P. R. Molecular evolution in nonrecombining regions of the *Drosophila melanogaster* genome. *Genome Biol. Evol.* **4**, 278–288 (2012).
26. Stolle, E. et al. Degenerative expansion of a young supergene. *Mol. Biol. Evol.* **36**, 553–561 (2018).
27. Gotzek, D., Clarke, J. & Shoemaker, D. Mitochondrial genome evolution in fire ants (Hymenoptera: Formicidae). *BMC Evol. Biol.* **10**, 300 (2010).
28. Pracana, R., Priyam, A., Levantis, I., Nichols, R. A. & Wurm, Y. The fire ant social chromosome supergene variant *Sb* shows low diversity but high divergence from *SB*. *Mol. Ecol.* **26**, 2864–2879 (2017).
29. Ross, K. G. & Shoemaker, D. Estimation of the number of founders of an invasive pest insect population: the fire ant *Solenopsis invicta* in the USA. *Proc. R. Soc. B* **275**, 2231–2240 (2008).
30. Ross, K. G., Vargo, E. L. & Keller, L. Social evolution in a new environment: the case of introduced fire ants. *Proc. Natl Acad. Sci. USA* **93**, 3021–3025 (1996).
31. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).
32. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118–128 (2005).
33. Huang, Y.-C. & Wang, J. Did the fire ant supergene evolve selfishly or socially? *Bioessays* **36**, 200–208 (2014).
34. Comeron, J. M., Williford, A. & Kliman, R. M. The Hill–Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* **100**, 19–31 (2008).
35. Kamdem, C., Fouet, C. & White, B. J. Chromosome arm-specific patterns of polymorphism associated with chromosomal inversions in the major African malaria vector, *Anopheles funestus*. *Mol. Ecol.* **26**, 5552–5566 (2017).
36. Jay, P. et al. Supergene evolution triggered by the introgression of a chromosomal inversion. *Curr. Biol.* **28**, 1839–1845 (2018).
37. Llaurens, V., Whibley, A. & Joron, M. Genetic architecture and balancing selection: the life and death of differentiated variants. *Mol. Ecol.* **26**, 2430–2448 (2017).
38. Tschinkel, W. R. *The Fire Ants* (Harvard Univ. Press, 2006).
39. Stevison, L. S., Hoehn, K. B. & Noor, M. A. F. Effects of inversions on within- and between-species recombination and divergence. *Genome Biol. Evol.* **3**, 830–841 (2011).
40. Crown, K. N., Miller, D. E., Sekelsky, J. & Hawley, R. S. Local inversion heterozygosity alters recombination throughout the genome. *Curr. Biol.* **28**, 2984–2990.e3 (2018).
41. Kelemen, R. K. & Vicoso, B. Complex history and differentiation patterns of the *t*-haplotype, a mouse meiotic driver. *Genetics* **208**, 365–375 (2018).
42. Grossen, C., Neuenschwander, S. & Perrin, N. The evolution of XY recombination: sexually antagonistic selection versus deleterious mutation load. *Evolution* **66**, 3155–3166 (2012).
43. Muller, H. J. The relation of recombination to mutational advance. *Mutat. Res. Mol. Mech. Mutagen.* **1**, 2–9 (1964).
44. Manfredini, F. et al. Molecular and social regulation of worker division of labour in fire ants. *Mol. Ecol.* **23**, 660–672 (2014).
45. Shoemaker, D. & Ascunce, M. S. A new method for distinguishing colony social forms of the fire ant, *Solenopsis invicta*. *J. Insect Sci.* **10**, 73 (2010).
46. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
47. Tang, H. et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).
48. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
49. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
50. Ometto, L., Shoemaker, D., Ross, K. G. & Keller, L. Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. *Mol. Biol. Evol.* **28**, 1381–1392 (2011).
51. Valles, S. M. & Porter, S. D. Identification of polygyne and monogyne fire ant colonies (*Solenopsis invicta*) by multiplex PCR of *Gp-9* alleles. *Insectes Soc.* **50**, 199–200 (2003).
52. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
53. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
54. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
55. Van der Auwera, G. A. et al. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinf.* **43**, 11.10.1–11.10.33 (2013).
56. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
57. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
58. Chen, K. et al. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. *Genome Res.* **24**, 310–317 (2014).
59. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
60. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
61. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at http://arXiv.org/abs/1207.3907 (2012).
62. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
63. Pitts, J. P., Camacho, G. P., Gotzek, D., Mchugh, J. V. & Ross, K. G. Revision of the fire ants of the *Solenopsis saevissima* species-group (Hymenoptera: Formicidae). *Proc. Entomol. Soc. Wash.* **120**, 308–411 (2018).
64. Lee, T.-H., Guo, H., Wang, X., Kim, C. & Paterson, A. H. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genom.* **15**, 162 (2014).
65. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
66. Reed, E. et al. A guide to genome-wide association analysis and post-analytic interrogation. *Stat. Med.* **34**, 3769–3792 (2015).
67. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
68. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
69. Moreau, C. S. & Bell, C. D. Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. *Evolution* **67**, 2240–2257 (2013).
70. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by somping trees. *BMC Evol. Biol.* **7**, 214 (2007).
71. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
72. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).
73. Xu, B. & Yang, Z. PAMLX: a graphical user interface for PAML. *Mol. Biol. Evol.* **30**, 2723–2724 (2013).


## Acknowledgements

We thank K. Harshman for Illumina and Pacbio sequencing support and R. Arguello, H. Darras, T. Flatt, J. Goudet, M. Qiaowei Pan, T. Schwander and J. Wang for comments on earlier versions of the manuscript. All computations were performed at the Vital-IT (http://www.vital-it.ch) Center for High-Performance Computing of the Swiss Institute of Bioinformatics. This work was supported by grants from the Swiss NSF to L.K., an ERC advanced grant to L.K., US NSF grants to K.G.R. and D.S. (no. 1354479) and K.G.R. and B.G.H. (no. 1755130) and US Federal Hatch funds to K.G.R. and B.G.H.

## Author contributions

D.G., K.G.R., D.S. and L.K. designed the experiments. K.G.R., D.S. and D.G. performed sample collection, DNA extraction and genotyping. Z.Y. performed PacBio sequence data collection and genome assembly and analysed the population genomic data. Z.Y., P.D., N.S. and L.K. conducted phylogenomic analyses. S.V.A., Z.Y., O.R.-G. and B.G.H. performed RNA-seq analyses. Q.H. conducted analyses of the structure of the genes interrupted by the inversions. S.H.M. performed population genetic simulations. Z.Y., K.G.R. and L.K. wrote the manuscript with the help of all authors.

## Competing interests

The authors declare no competing interests.

## Additional information

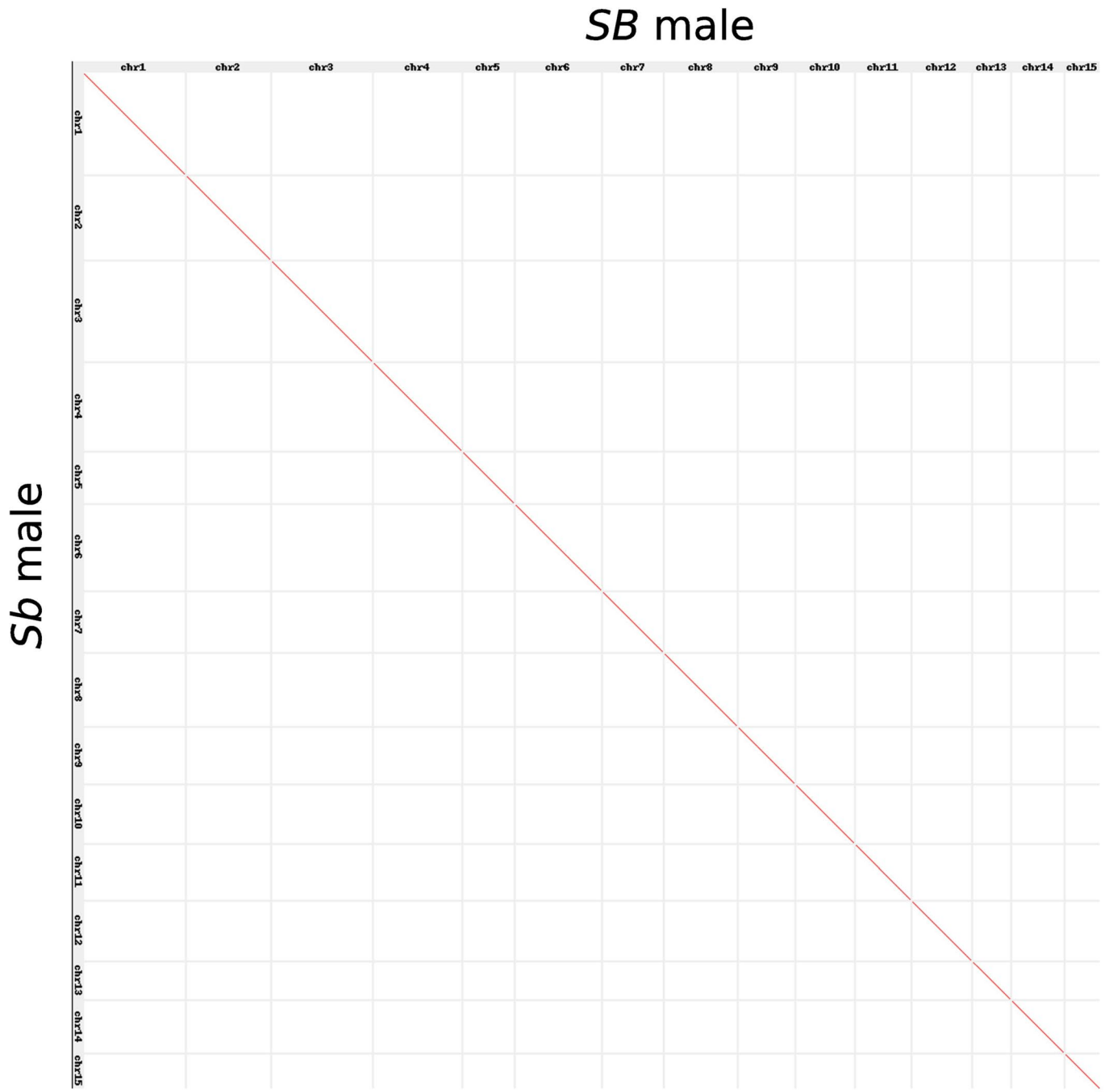
**Extended data** is available for this paper at https://doi.org/10.1038/s41559-019-1081-1.

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41559-019-1081-1.

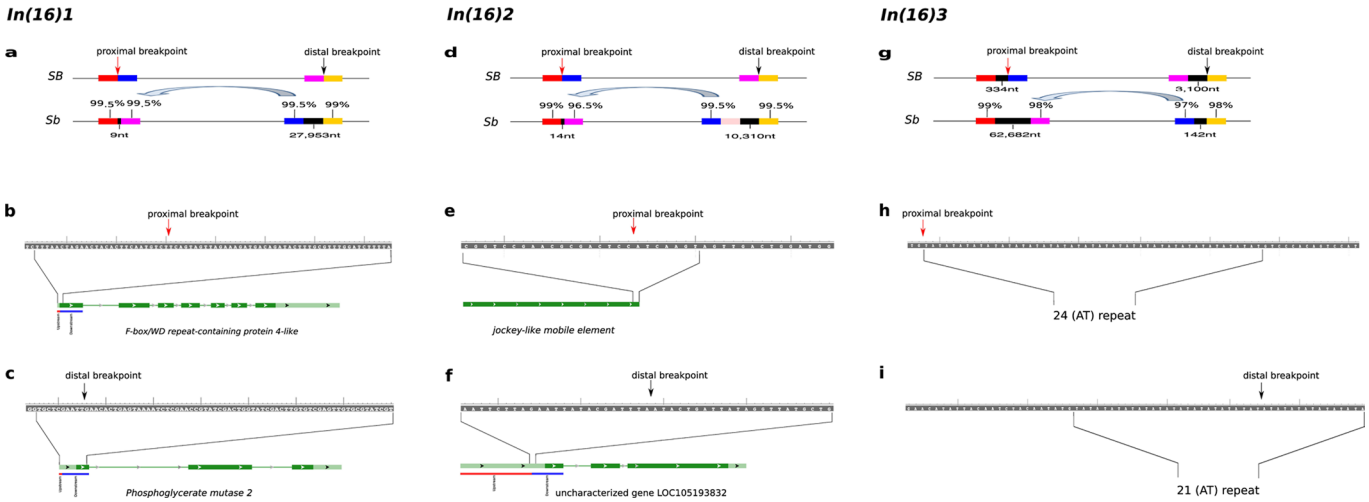
**Correspondence and requests for materials** should be addressed to K.G.R. or L.K.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.




**Extended Data Fig. 1 | Dot plot of the alignment between chromosomes 1-15 of *Sb* and *SB* males of *S. invicta* from the invasive (United States) range.** The red line indicates the forward strand alignment.



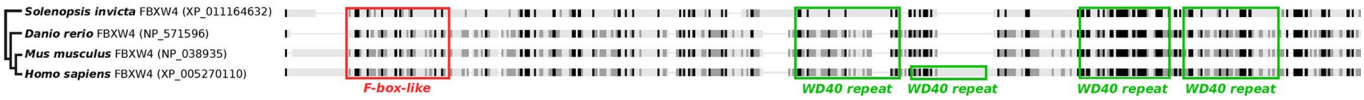

**Extended Data Fig. 2 | Sequence features for the breakpoints of the three *Sb* supergene inversions, *In(16)1* [chr16:14,549,064-24,031,576], *In(16)2* [chr16: 13,705,210-24,030,990], and *In(16)3* [chr16: 12,612,565-13,683,100]. a**, The red and blue blocks represent the 200nt segments adjacent to the *In(16)1* proximal breakpoint (red arrow), and the magenta and gold blocks represent the 200nt segments adjacent to the distal breakpoint (black arrow) of this inversion. The blue and magenta blocks, and the segment between them, are inverted between *SB* and *Sb*, as indicated by the grey arrow. The *Sb* haplotype has a 9nt insertion at the proximal breakpoint and a 27,953nt insertion at the distal breakpoint (black blocks). Percentages are the sequence similarity (disregarding deletions) between *SB* and *Sb* for the 200nt segments immediately upstream and downstream of the breakpoints. **b**, The *In(16)1* proximal breakpoint (red arrow) on the *SB* haplotype is located in exon 1 of the "*F-box/WD repeat-containing protein 4-like*" gene (NCBI Gene symbol: LOC105199310; green blocks depict exons; pale green represents the UTRs and dark green the coding sequence [CDS] regions). The red and blue lines under exon 1 indicate the segments that are upstream or downstream of the proximal breakpoint. **c**, The *In(16)1* distal breakpoint (black arrow) on the *SB* haplotype is located in the 5′ UTR of the "*Phosphoglycerate mutase 2*" gene (NCBI Gene symbol: LOC105193833; green blocks depict exons; pale green represents the UTRs and dark green the CDS regions). The red and blue lines under exon 1 indicate the segments that are upstream or downstream of the distal breakpoint. **d**, The red and blue blocks represent the 200nt segments adjacent to the *In(16)2* proximal breakpoint (red arrow), and the magenta and gold blocks the 200nt segments adjacent to the distal breakpoint (black arrow) of this inversion. The red block contains the 3′ end of a Jockey-like mobile element. The *Sb* haplotype has a 14nt insertion (black block) at the proximal breakpoint as well as a second Jockey-like mobile element gene (pink block) and a 10,310nt insertion (black block) just upstream of the distal breakpoint. **e**, The *In(16)2* proximal breakpoint (red arrow) in the *SB* haplotype is located in the single exon (dark green) of a Jockey-like mobile element. **f**, The *In(16)2* distal breakpoint (black arrow) in the *SB* haplotype is located in the 5′ UTR of the uncharacterized gene "LOC105193832" (containing 3 exons depicted as green blocks; pale green represents the UTR and dark green the CDS region). **g**, The red and blue blocks represent the 200nt segments adjacent to the *In(16)3* proximal breakpoint (red arrow), and the magenta and gold blocks the 200nt segments adjacent to the distal breakpoint (black arrow) of this inversion. The *SB* haplotype has a 334nt insertion at the proximal breakpoint and a 3,100nt insertion at the distal breakpoint (black blocks); both are absent in the *Sb* haplotype, which instead has a 62,682nt insertion at the proximal breakpoint and a 142nt insertion at the distal breakpoint (black blocks). **h**, The *In(16)3* proximal breakpoint (red arrow) in the *SB* haplotype is located just upstream of a region containing 24 dinucleotide (AT) repeats. **i**, The *In(16)3* distal breakpoint (black arrow) in the *SB* haplotype is located within a region containing 21 dinucleotide (AT) repeats.

**a**

***Phosphoglycerate mutase 2* (*PGAM2*)**


**b**

***F-box/WD repeat-containing protein 4-like* (*FBXW4*)**



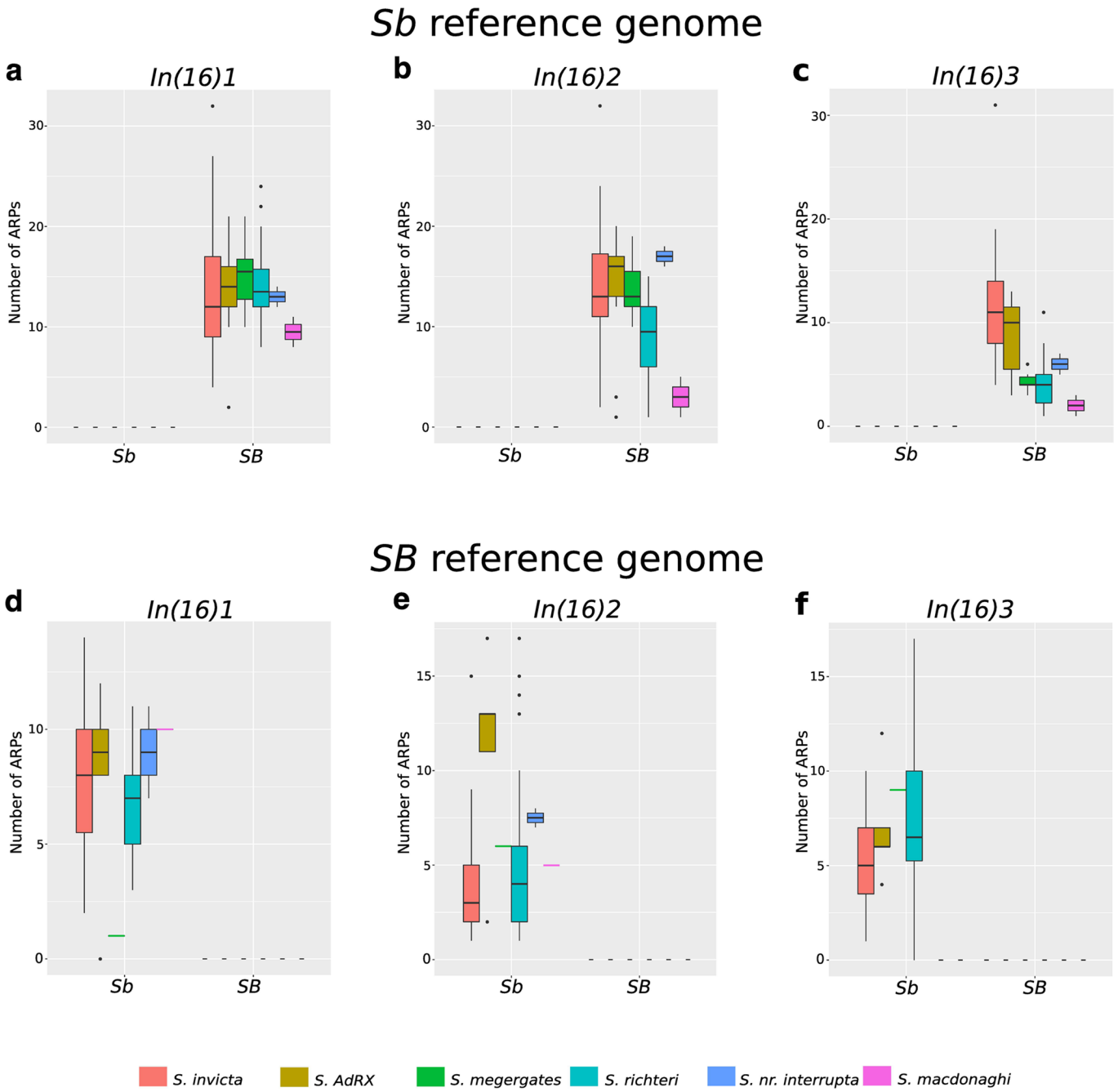

**c**

| *Solenopsis invicta* | | *Drosophila melanogaster* | Viability Effects of Gene Disruption (Flybase) | Substrates |
|---|---|---|---|---|
| PGAM2 (XP_011156752) | *ortholog** | PGAM (CG1721) | Viable | |
| FBXW4 (XP_011164632) | *paralogs*** | *Ago (CG15010)* | Lethal | Trh, CycE, dMyc, Notch |
| | | *Slimb (CG3412)* | Most are lethal | ARM, Ci, Cact, Dl, E2F, PER, PLK4, Rel |
| | | *Fbw5 (CG9144)* | Viable | Myc, trh, sima, CycE |

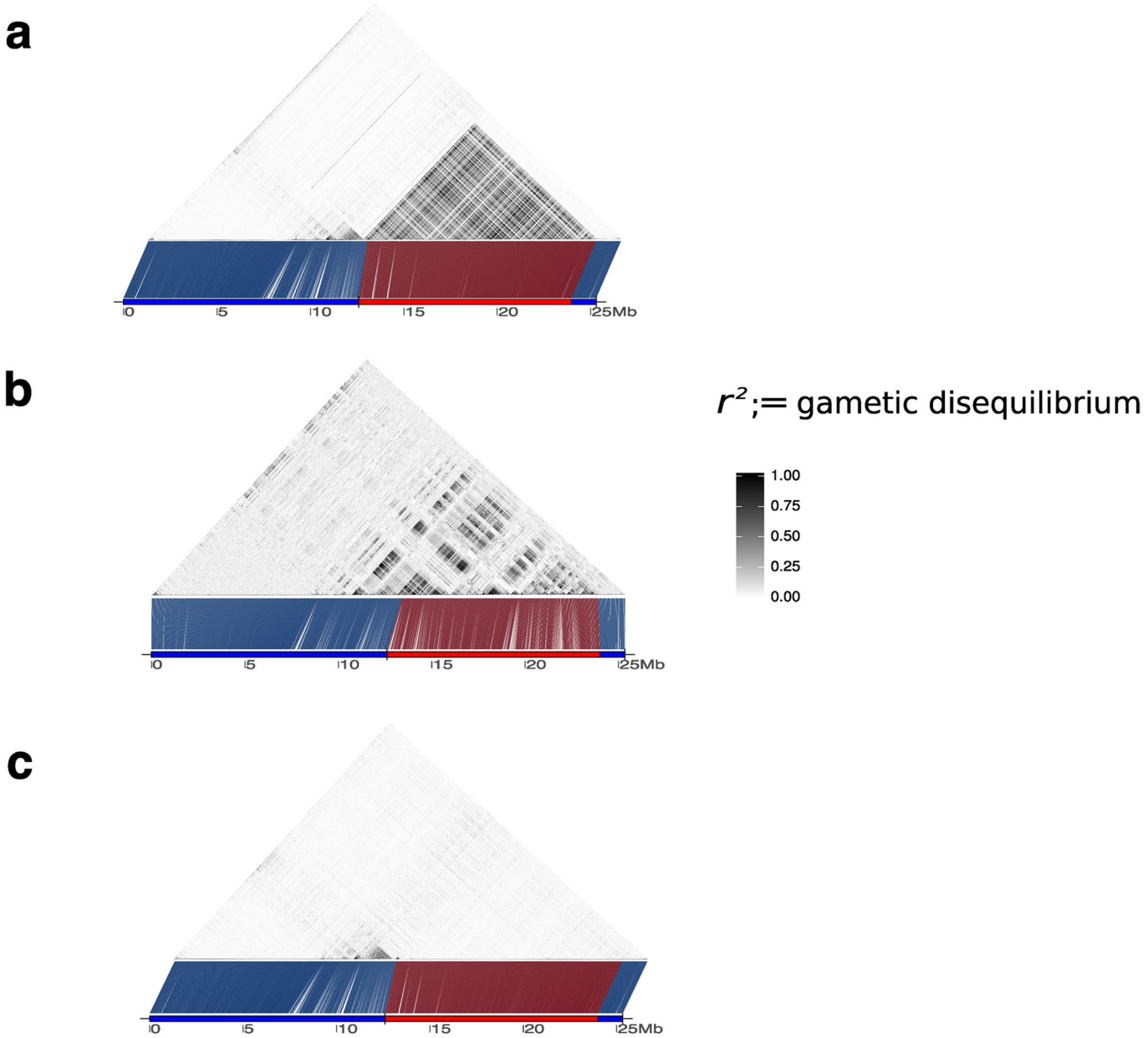
* Based on reciprocal best BLAST

**Based on the protein domains combination

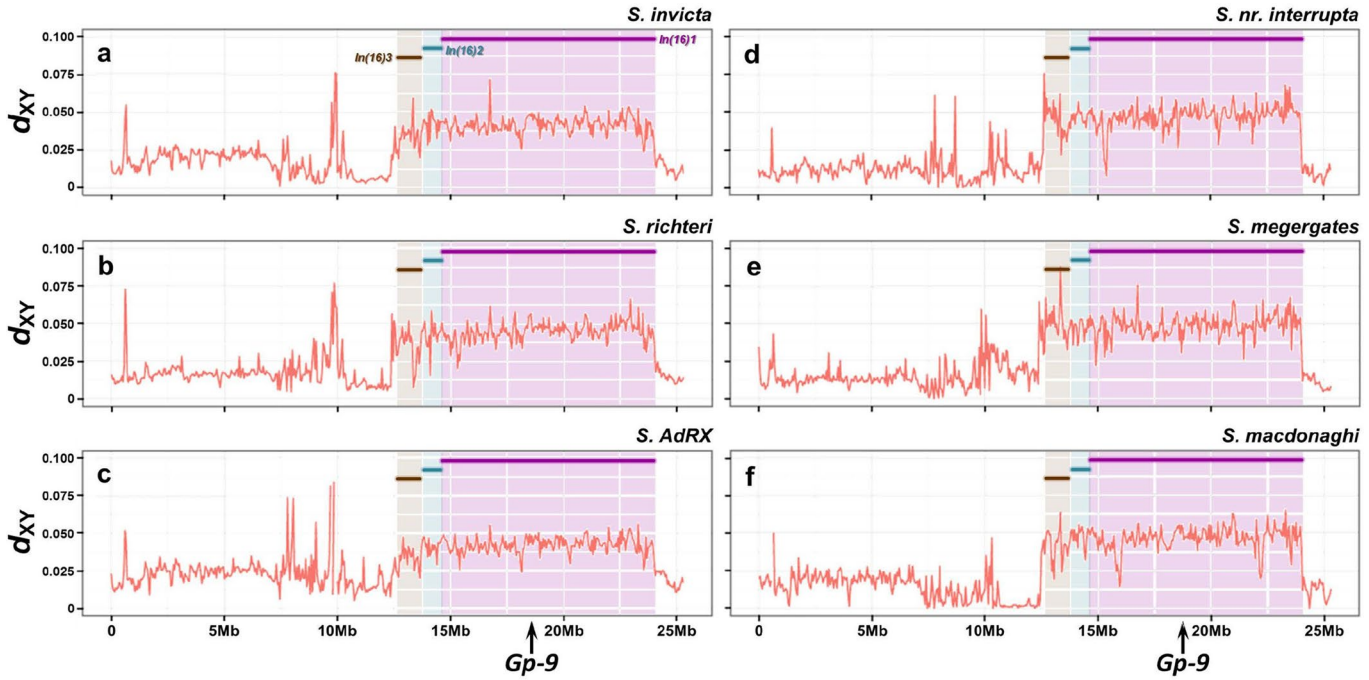
**Extended Data Fig. 3 | Functional features of the two protein-coding genes interrupted by the inversion *In(16)1* breakpoints. a**, The *S. invicta Phosphoglycerate mutase 2* (*PGAM2*) protein sequence shows a very high level of similarity with putative orthologs in *Drosophila melanogaster* (fruit fly), *Danio rerio* (zebrafish), *Mus musculus* (mouse), and *Homo sapiens*. Virtually all described functional sites are conserved across the five species: the catalytic core is strictly conserved and all but one amino acid of the substrate binding sites are also conserved. This high level of amino acid sequence conservation suggests conservation of the function of *PGAM2* among putative orthologs. **b**, The *S. invicta FBXW4* protein contains the typical domains of F-box protein ubiquitin ligase complexes (F-box and WD40 repeats from InterPro and NCBI predictions) and is therefore also likely to be involved in ubiquitination and proteasome degradation. In the alignment, identical sites are shown with black bars, 75% similar with dark grey, 50% with light grey, and 25% with white. **c:** Disruption of *PGAM* in *D. melanogaster*, the ortholog of *PGAM2* in *S. invicta*, is not lethal. Paralogs of *S. invicta FBXW4* in *D. melanogaster* have a wide range of substrates and their disruption can be lethal.



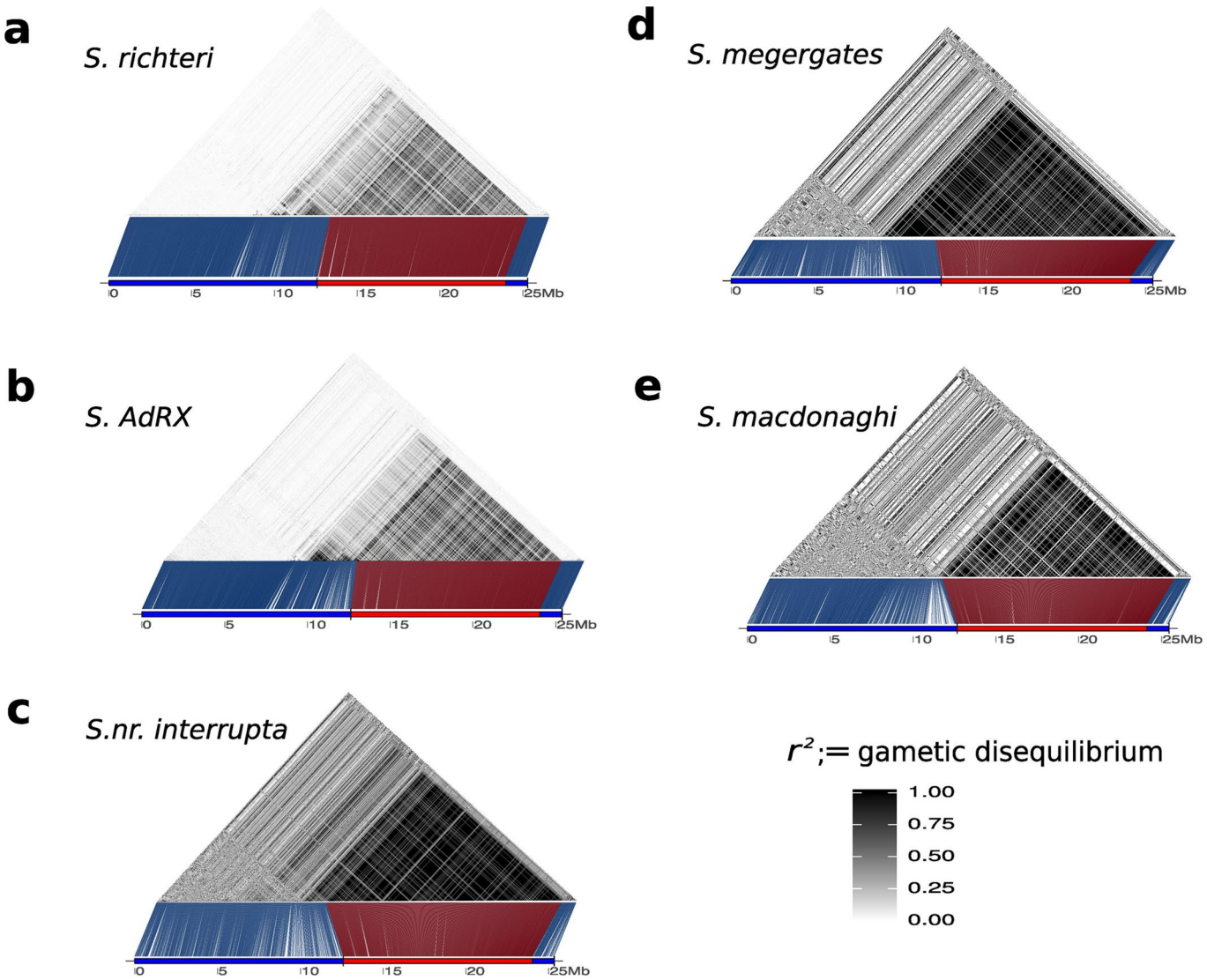

**Extended Data Fig. 4 | The box plots of the numbers of anomalous read pairs (ARPs) connecting the downstream and upstream regions (400nt) adjacent to the breakpoints of the three supergene inversions [*In(16)1*, *In(16)2*, *In(16)3*] for *Sb* and *SB* males of six socially polymorphic fire ant species.** Each box ranges from the first (Q1) to the third quartile (Q3) of the distribution and represents the interquartile range (IQR). A line across the box indicates the median. The whiskers are lines extending from Q1 and Q3 to end points that are defined as the most extreme data points within Q1 − 1.5 × IQR and Q3 + 1.5 × IQR, respectively. Each outlier outside the whiskers is represented by a solid dot. **a**-**c**, ARPs connecting proximal and distal inversion breakpoints when samples are mapped to the *Sb* reference genome. **d**-**f**, ARPs connecting proximal and distal inversion breakpoints when samples are mapped to the *SB* reference genome. There are four *Gp-9*$^{b}$ individuals with zero values when samples are mapped on the *SB* reference genome (one S. *AdRX* for inversion *In(16)1)*; one *S. richteri*, one *S. nr. interrupta*, and one *S. macdonaghi* for inversion *In(16)3*)). However, the targeted local assembly of the breakpoint sequences yielded contigs that bridge these breakpoints in *Sb* males of all these four individuals.



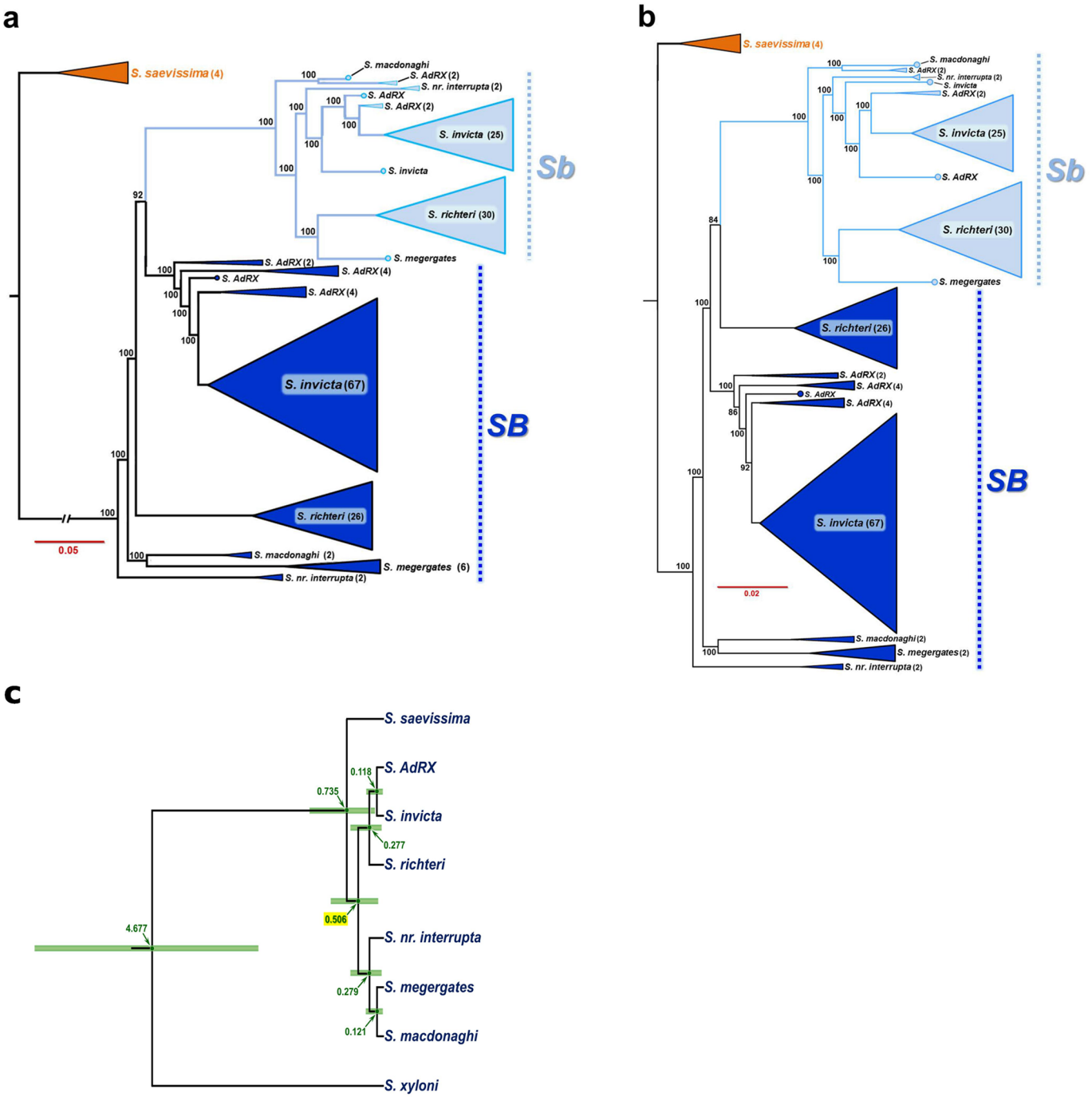

**Extended Data Fig. 5 | Linkage disequilibrium ($r^2$;= gametic disequilibrium for haploid males) in native *S. invicta* estimated using SNPs across the social chromosome (chr16). a**, LD dot plot for pooled *SB* (N = 60) and *Sb* (N = 19) males. **b**, LD plot for *Sb* males. **c:** LD plot for *SB* males. The coloured bar under each plot represents the physical map of the chromosome, with the red segment indicating the region where recombination is suppressed between the *Sb* and *SB* haplotypes in invasive (United States) *S. invicta*. SNPs are ordered according to physical position on the chromosome. The blue and red dashed lines link SNPs on the LD plot to their position on the physical map. The centromere occupies the approximate region 7.5-11 Mb.



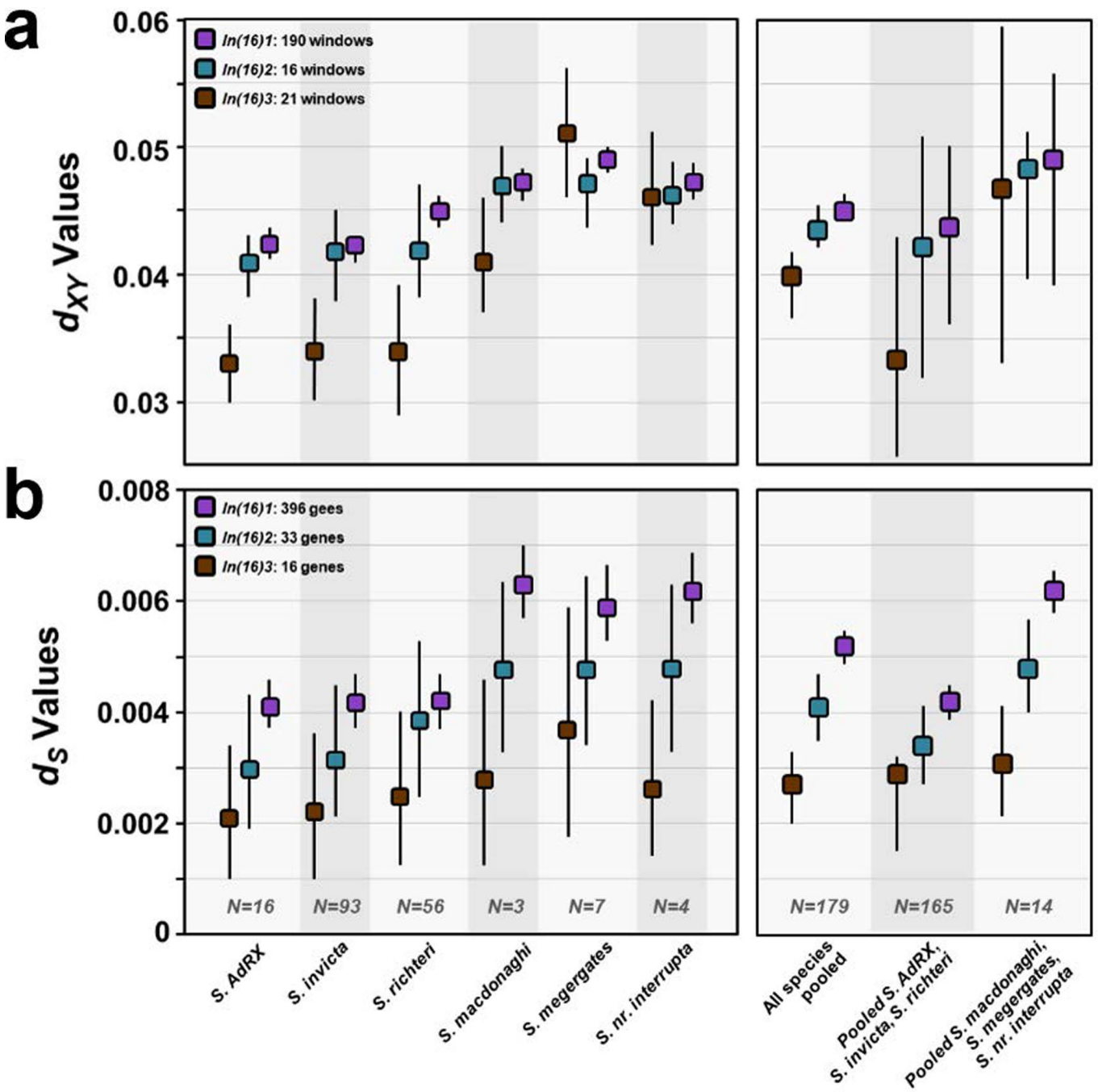

**Extended Data Fig. 6 | Nucleotide divergence ($d_{XY}$) values between sequences from conspecific *Sb* and *SB* males of the six socially polymorphic fire ant species estimated using 50 kb non-overlapping sliding windows across the social chromosome.** The boundaries of the differently shaded intervals correspond to the breakpoints of the three inversions, with the inversions depicted by thick horizontal coloured lines. The x axes represent physical position along chromosome 16, with the location of gene *Gp-9* indicated. For each of the six species, $d_{XY}$ values between *Sb* and *SB* males were significantly higher for the region corresponding to the inversions than for the rest of the social chromosome (Mann–Whitney U-tests, all $P < 0.01$). The region of elevated $d_{XY}$ proximal to the inversions (at ca. 8-10 Mb) appears to experience reduced recombination in *S. invicta* based on progeny studies;[15] this area may correspond at least partly to the centomeric region or some other, unknown feature may be responsible for the reduced recombination and elevated $d_{XY}$ there.




**Extended Data Fig. 7 | Linkage disequilibrium ($r^2$; = gametic disequilibrium) for pooled conspecific *Sb* and *SB* males of five native populations of socially polymorphic fire ants estimated using SNPs across the social chromosome. a**, LD dot plot for pooled *Sb* (N = 30) and *SB* (N = 26) *S. richteri* males. **b**:, LD plot for *Sb* (N = 5) and *SB* (N = 11) *S. AdRX* males. **c**, LD plot for *Sb* (N = 2) and *SB* (N = 2) *S. nr. interrupta* males. **d**, LD plot for *Sb* (N = 1) and *SB* (N = 2) *S. megergates* males. **e**, LD plot for *Sb* (N = 1) and *SB* (N = 2) *S. macdonaghi* males. Mean LD estimates for exclusively *Sb* haplotypes are $r^2 = 0.41$, 0.36, 0.85, 0.87, and 0.80 for *S. richteri, S. AdRX, S. megergates, S. nr. interrupta*, and *S. macdonaghi*, respectively. See Extended Data Fig. 5 caption for additional information.




**Extended Data Fig. 8 | Phylogenetic trees for the supergene region of the social chromosome and for the fire ant species included in this study.** **a**, Alternative maximum-likelihood (ML) tree for the supergene region that disregards the LD (non-independence) of the 30,921 included SNPs. **b**, Alternative ML tree for the supergene region that accounts for LD, using a pruning threshold of 0.5. Trees in (**a**) and (**b**) were rooted with the outgroup species *S. saevissima*, which lacks the three chr16 inversions. The red scale bars are substitutions per site. **c**, Bayesian-inference tree with divergence time estimates for the study species based on sequences of five nuclear genes. The analysis incorporated the uncorrelated molecular clock method (BEAST), with the age of the basal divergence calibrated using data from a previous study (see Methods). The tree was rooted with the outgroup species *S. xyloni*. The number at each node represents the mean estimate of divergence time (in Myr), with the green bars representing the 95% confidence intervals (CIs) about the estimates; divergence time for the two major lineages of the socially polymorphic clade is highlighted with yellow background. Note that the topology of this tree is fully congruent with that of the ML species tree based on 12,237,341 non-supergene SNPs (Fig. 4).




**Extended Data Fig. 9 | Evidence from $d_{XY}$ and $d_S$ estimates that the three inversions comprising the *Sb* haplotype emerged in the order *In(16)1 → In(16)2 → In(16)3*.** **a**, Values of $d_{XY}$ between conspecific *Sb* and *SB* haplotypes for the three inversions in each of the six socially polymorphic fire ant species studied (left panel) and for groups of species for which data were pooled (right panel) (means and 95% CIs from 1000 bootstrap replicates). Values for subsets of species were pooled if they did not differ significantly (see Methods). Note that $d_{XY}$ values for *S. macdonaghi*, *S. megergates*, and *S. nr. interrupta* may not be highly accurate because of the small sample sizes for *Sb* in these species. **b**, Values of $d_S$ between conspecific *Sb* and *SB* haplotypes for the three inversions in each of the six socially polymorphic fire ant species studied (left panel) and for groups of species for which data were pooled (right panel) (means and 95% CIs from 1000 bootstrap replicates). Values for subsets of species were pooled if they did not differ significantly (see Methods). Note that $d_S$ values for *S. macdonaghi*, *S. megergates*, and *S. nr. interrupta* may be inflated because the small sample sizes of *Sb* for these species mean that some intrahaplotype polymorphic synonymous changes are interpreted to be fixed synonymous differences.

nature research

Corresponding author(s): Laurent Keller

Last updated by author(s): Nov 9, 2019

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No software were used for data collection. |
| Data analysis | Open source software were used in the study, including BWA, Bamtools, Freebayes, BEAST, Canu etc. Detailed information listed in Methods section. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The genome assembly, gene models, and sequence reads are available at the NCBI under the BioProject PRJNA421367.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | We conduct comparative whole-genome sequence analyses of 169 samples of 7 fire ant species to explore the origin of a large supergene region that undergoes limited recombination and underlies important variation in colony social organization. |
| Research sample | The native fire ants (Solenopsis spp.) showing variation in social organisation were collected for this study. |
| Sampling strategy | We used haploid males exclusively for genome sequencing to directly infer haplotypes. We used only a single male from each monogyne colony in order to avoid sequencing related individuals and, whenever possible, we sequenced one Sb and one SB male from each polygyne colony. |
| Data collection | Genome DNA were extracted from male ants. Standard Illumina protocols (TruSeq DNA) were used to prepare the paired-end libraries. Each genome was sequenced to at least 10X sequencing depth. Zheng Yan collected the raw sequence reads from Sequence platform. |
| Timing and spatial scale | Samples were collected from the native ranges of each fire ant species in South America during collection trips from 1990 to 2015. Description of research ants used for experiments can be found in the Table S1. |
| Data exclusions | No data exclusions were performed for this study. |
| Reproducibility | Replicate data analysis were successful. |
| Randomization | No randomization of ants. Ants analyzed were genotyped and social form-matched whenever possible. |
| Blinding | Investigators were not blinded to ant genotypes during experiments. Result reported for ant genomic data are not subjective but rather based on the extensive genotype and DNA sequence. |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☐ | ☒ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

| | |
|---|---|
| Laboratory animals | No laboratory animal was used in the study. |
| Wild animals | *Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.* |
| Field-collected samples | *For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.* |
| Ethics oversight | *Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.* |

Note that full information on the approval of the study protocol must also be provided in the manuscript.