The molecular basis of complex phenotypes in the fire ant *Solenopsis invicta*

Carlos Martínez Ruiz

September 2019

Queen Mary
University of London

Thesis submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy

I,Carlos Martínez Ruiz, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date:

Details of collaboration and publications:

All chapters in the present thesis were written by me and reviewed by my supervisors and/or colleagues.

Chapter 2 collaborators: Dr. Rodrigo Pracana, Dr. Eckart Stolle and Dr. Monika Struebig

Chapter 4 collaborators: Dr. Marc Robinson-Réchavi, Dr. Ed Vargo and Marian Priebe

More details about the specific contributions of each collaborator are available at the beginning of each chapter.

General abstract

Understanding the phenotypic diversity we observe across the tree of life is a fundamental part of research in biology. A particular case of special interest is the occurrence of several discrete phenotypes within the same population. These cases allow us to interrogate the basis of phenotypic differences without the influence of confounding factors such as phylogenetic distance or geographical barriers. The steady reduction in sequencing price of the past decades has resulted in more molecular data available for specific organisms where phenotypic differences occur within a single population.

This thesis focuses on one such organism, the fire ant *Solenopsis invicta*, to shed light on the molecular basis of within population discrete phenotypic diversity. This species displays two types of such differences. It is socially polymorphic, which means that colonies can have two types of social organisation. Additionally, *S. invicta* has three types of morphological castes: queens, workers and males. Social polymorphism in this species is determined genetically, whereas caste differences are triggered by environmental cues during development. *S. invicta* is thus a good study system to understand different molecular mechanisms underlying discrete phenotypes. In this thesis, I use a combination of molecular data and modelling approaches to investigate the molecular underpinning of both types of phenotypic differences in *S. invicta*.

I show that the genetic architecture maintaining different types of social organisation is likely to have arisen by a combination of evolutionary conflict and a strong influence of gene flow fluctuations between phenotypes. I also show that caste differences depend on the expression of thousands of genes, most of which are tissue specific. These results are put in the context of the wider theoretical framework, and provide further understanding on the evolution of phenotypic diversity.

Acknowledgements

This thesis would not have been possible without the support, financial and otherwise, of many people and institutions throughout these last four years, but also throughout my life.
First of all, I would like to thank my funders, the London NERC DTP (grant number NE/L002485/1) to provide both the financial and academic support of this project.
I am extremely grateful to my supervisors, Prof. Richard Nichols and Dr. Yannick Wurm for their constant support and insight into my research. Their input has made me grow as a person and as a scientist.
The work presented here would not have been possible without the collaborations with other researchers and institutions. Dr. Marc Robinson-Réchavi, from the University of Lausanne (Switzerland) has provided financial support and advice to generate the most comprehensive tissue-specific RNAseq dataset for hymenopterans to date. Dr. Ed Vargo, Dr. Elly Espinoza and Dr. Laura Johnson from Texas A&M University for their help with the field collections of fire ant colonies.
From Queen Mary University, for their technical support, I would like to thank Phil Howard, Dr. Monika Stuerbig, Dr. Chloe Economou and Martin Tran for their patience with me in the wet lab (and keeping me safe from my own mistakes!). This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. I therefore, I also need to mention the invaluable work that the IT technicians in SBCS, Adrian Lärkeryd and James Crowe have done and do to keep our jobs running smoothly in the HPC.
My work at Queen Mary would have been much harder if it weren't for my colleagues at the Wurm Lab, past and present: Dr. Eckart Stolle, Dr. Joe Colgan, Bruno Vieira, Dr. Leandro Rodrigues Santiago, Dr. Federico López-Osorio, Raphaella Jackson, Dr. Anindita Brahma, Gabriel Hernandez-Gomez and Marian Priebe. But I want to thank especially Dr. Rodrigo Pracana, Anurag Priyam, Magda Shacht and Emeline Favreau, for always being there for me, for sharing despair, laughs and/or a drink. Outside of the Wurmlab I would like to thank my colleagues at the department of Organismal Biology, who have shaped me and my work throughout these years. I would like to thank especially Sandra Álvarez-Carretero, Giacomo Vitali, Emma Lockley, Andy Knapp, Dan Nicholson and Dr. Jasmin Zohren for the laughs, the science and just for being all round great human beings. Here I would like to include my colleagues from the London NERC DTP cohort 2, with who I shared the first 6 months of this PhD program, who have been a much needed support network during the duration of my project.
And for keeping me sane during the last two years, put up with me even when I didn't stand myself and being probably one of the kindest, smartest, funniest people to have ever existed: thank you Sarah. I cannot put down in words how lucky I am to have you in my life.
Por último, no puedo dejar de agradecer a la gente que me ha acompañado y dado apoyo durante todos estos años, y que me han llevado, en última instancia, a acabar este proyecto. Gracias a los magníficos profesores del IES Riba-Roja de Túria, por su generosidad y paciencia, por enseñarme los primeros pasos de la vida profesional. Gracias a Dr. Edu García Roger y Dr. Karmen Rojo, por darme mi primera oportunidad en el mundo investigador, y que, estoy seguro, me han abierto las puertas a esta aventura. Y por último, y, por ello, más importante, gracias a mis padres, Dolo y Carlos, por la vida que me habéis dado. Por haber estado siempre ahí para mi, incluso (especialmente) cuando os lo he puesto difícil. Nada de esto sería posible sin vosotros, gracias de corazón.

To the ants that gave their lives to make this research possible

Table of Contents

Chapter 1: Introduction

Bridging the gap between evolutionary theory and molecular data

The current theoretical framework explaining adaptation is largely based on modern evolutionary synthesis from the mid-twentieth century. In this framework, the determination of a phenotype is viewed as a unidirectional process, from genes to characters. However, data suggests that this relationship is most likely bidirectional, with phenotypes shaping the structure of genotype and vice versa (Nevo 2001). Although these facts have been known for long and some effort has been put in incorporating new findings to theory (reviewed in Orr 2005), most of the models used for explaining adaptation rely on the old assumptions of the modern synthesis such as the independent spread of genes or the lack of non-additive genetic interactions (Mayr 1982). The reason for the widespread use of these simplifying assumptions is that they have allowed the development of an extraordinary mathematical body which is well understood by geneticists and which explains part of data observed (Orr 2005). As a result, the molecular basis of phenotypic diversity is not entirely understood, and new approaches are needed to bridge the gap between the recent molecular data and evolutionary theory.

The drop in sequencing prices during the last decades has resulted in an explosion of genome-wide molecular data for a wide range of organisms. Data that extends beyond the small number of traditional genetic model species that had previously been studied in detail. This new ability to obtain data open up opportunities to study the interface between genotypes and phenotypes. Of particular interest are the organisms displaying within-population stable complex polymorphisms. This concept refers to instances where two or more discrete complex phenotypes co-exist within a population of the same species. The term “complex” is used here to refer to the fact that these phenotypes differ in several traits, so that they are likely to be encoded by the interaction of multiple genes. Such stable complex polymorphisms are common across the tree of life. Some are widespread, such as sexual dimorphism, others are more species-specific, for instance, multiple male morphs in some dung beetles (Matsumoto & Knell 2017), neotenic morphs in tiger salamanders (Lackey et al. 2019) or swarming behaviour in locusts (Ayali 2019). The different phenotypes may be directly encoded in specific regions of the genome, as happens with sex chromosomes responsible for triggering sexual dimorphism. On the other hand, other stable

polymorphisms are controlled by environmental factors, and the mechanisms at the molecular level involve changes in gene expression, as is the case with the swarming behaviour of locusts (Kang et al. 2004).

Stable complex polymorphisms provide valuable study systems for understanding the interactions between phenotypes and genotypes: they allow the study of discrete trait differences, but because they occur within the same population, the molecular differences underlying different phenotypes are not confounded with factors such as geography or taxonomy. In this thesis, I use the red fire ant *Solenopsis invicta* as a model to shed light on the molecular processes underlying stable complex polymorphisms. This species displays two types of stable polymorphism. At the colony level, it is socially polymorphic: fire ants can display two types of social organisation. Colonies can either have one or multiple queens per colony. This social polymorphism is determined genetically, through a low recombination region of the genome designated as a supergene (Wang et al. 2013). At an individual level, as with other social insect species, the fire ants have three morphological castes, queens, workers and males. These castes represent three discrete phenotypes at the morphological, behavioural and reproductive level. Unlike social polymorphism, caste differences are determined environmentally (Tschinkel 2006). *S. invicta* thus provides an example of complex stable polymorphisms maintained by two types of molecular regulation.

The following sections will introduce in more detail the nature of supergenes, their role in evolution, the mechanisms of caste determination in social insects, the life history of *S. invicta* and how these topics will be addressed in the present thesis.

The *Solenopsis invicta* system

S. invicta has been well studied, mainly due to its pest status as an invasive species in the USA since its introduction from South America in the 1930s (Callcott and Collins, 1996). More recently it has spread elsewhere in the world, including Australia, Taiwan and Southern China (Ascunce et al. 2011). It is an example of social polymorphism and, additionally *S. invicta* is relatively easy to rear in the laboratory (Tschinkel 2006), which, added to the fact that its genome sequence was published in 2011 (Wurm et al. 2011), has led to *S. invicta* emerging as a model organism for the study of social evolution. This role has in turn boosted the production of data available for this species.

S. invicta colonies have either a single or multiple queens (Tschinkel 2006). This apparently simple phenotypic trait affects the whole organisation, behaviour and physiology of the whole colony. Single-queen colonies have bigger queens and workers (Greenberg et al. 1985),

moreover their queens are more fertile, more efficient egg layers, disperse over longer distances and have a longer lifespan than their multiple-queen counterparts (DeHeer 2002). After dispersal, queens destined to have their own single-queen colony are able to disperse over large distances and independently form new colonies. On the other hand, queens destined to be part of multiple-queen colonies can either become reproductive in their natal nests or found new colonies, although they are not able to do so on their own. These queens need the assistance of workers from the natal colony to successfully establish a new nest, usually by budding from the original colony (Vargo & Porter 1989). At the colony level, workers from single-queen colonies are more territorial and aggressive than those from multiple-queen colonies (Chirino et al. 2012). Single-queen workers only accept one reproductive queen, killing any supernumerary pretenders to the crown, whereas workers in multiple-queen colonies are far less aggressive towards new reproductive queens (DeHeer 2002). Differences in colony social form result in important changes in the balance of fitness costs within a colony. Because workers in multiple-queen colonies are less related to the queens in the colony than in single-queen colonies, the fitness cost of altruistic behaviour for workers is likely to be different in either colony type (Keller 1995).

Intriguingly, all these phenotypic features of the colony are controlled by a single Mendelian factor (Ross & Shoemaker 1997), the two alleles of the gene Gp-9 (Krieger & Ross 2002). At first it was considered that this gene could have large down-stream phenotypic effects that would account for the striking differences observed in the *S. invicta* colonies (Gotzek & Ross 2007). However it was later demonstrated that this gene is actually part of a larger non-recombining region -supergene- which would actually define the two social phenotypes (Wang et al. 2013). This supergene, which has been described as a ‘social chromosome’, is 20.9 Mb long (Stolle et al. 2019), including around 400 genes and taking up to 55% of an actual chromosome (Pracana et al. 2017). The two variants of this social chromosome are referred to as SB and Sb. All individuals in single-queen colonies are homozygotes SB/SB, thus producing diploid SB/SB reproductive females and SB haploid reproductive males (Wang et al. 2013). The association with the social chromosome with multiple-queen colonies is less straightforward. All queens from multiple-queen colonies are SB/Sb, any SB/SB young reproductive female is systematically killed by workers (Keller & Ross 1998), although sometimes a few are able to escape (DeHeer 2002). Workers, however, can either be SB/SB or SB/Sb, it has been shown that only 5%-15% of the workers need to be SB/Sb heterozygotes to form a multiple-queen colony (Ross & Keller 2002). In wild populations, the proportion of homozygote workers is estimated to be between 20% to 50% (Buechel et al. 2014). Multiple-queen colonies thus produce mostly SB/Sb and very few SB/SB queens and both SB and Sb haploid males (Gotzek & Ross 2007). Sb/Sb queens have been very rarely

reported in mating flights and inside some nests (Fritz et al. 2006), but there is no evidence for their offspring to be viable, and are thus considered a lethal recessive and an evolutionary dead end (Tschinkel 2006).

In the invasive populations in USA, multiple-queen colonies produce large proportions (up to 90%) of diploid males which are mostly sterile. This has been interpreted as a result of the loss of genetic diversity in invasive populations (Ross et al. 1993). In native populations of *S. invicta* diploid males are less common (around 14%), but they still are an exclusive feature of multiple-queen colonies (Ross et al. 1993). Diploid males are very costly for the colony, do not contribute to colony maintenance and the vast majority of them are sterile or produce unviable triploid offspring (Ross & Fletcher 1986). In the vulnerable early stages of colony founding, single-queen colonies where a half of the individuals are diploid males, the colony would be very unlikely to survive because of these costs. Established single-queen colonies almost never produce diploid males. In multiple-queen colonies, however, given that there are several queens contributing to producing individuals, the cost of diploid males appears to be bearable and the colony survives (Ross & Fletcher 1986).

In addition to the production of diploid males, haploid males from multiple-queen colonies disperse over shorter distances than those from single-queen colonies. As a consequence most multiple-queen successful reproductive females would be mating with SB haploid males from single-queen colonies (Ross & Keller 1995b). Finally, males from multiple-queen colonies carrying the Sb allele produce less sperm and the females they mate with are far more likely to re-mate than those fertilised by SB males (Lawson et al. 2012). According to some authors, these factors would result in an uneven, unidirectional gene flow between social forms (Ross & Shoemaker 1993), where most successful reproductives tend to be produced by single-queen colonies. Other studies, however, suggest that the gene flow is limited between social forms (Goodisman et al. 2000). This limited gene flow would be achieved by biasing the production of reproductive individuals towards Sb carrying males (Fritz et al. 2006) and queens (Keller & Ross 1998; Buechel et al. 2014) in multiple queen colonies. Because the Sb variant is only present in multiple-queen colonies, most reproductive individuals from these colonies would only produce colonies from the same social form, thus limiting the gene flow between phenotypes. Additionally, other mechanisms such as selective mating within social form may counteract the lower reproductive success of multiple-queen individuals (Saddoris et al. 2016). It is also relevant to note that the idea of a unidirectional gene flow is based on the assumption that multiple-queen colonies produce a high proportion of sterile diploid males. This assumption is true for populations in the invasive range of North America, but not for the native range (Ross et al. 1993).

An understanding of the life history of *Solenopsis invicta* is required to unravel the evolutionary forces at play in the emergence of a stable social polymorphism in this species. But life history is, however, just part of the story. To produce a full picture, the molecular basis needs to be understood too. The next section will deal with such mechanisms: the supergenes. What are they? What are the evolutionary forces leading to their emergence? And finally, how do they define and maintain different phenotypes in *S. invicta*?

Supergenes and their role in evolution

Supergenes are structures in the genome that contain tightly linked genes which allow the maintenance of complex genetic interactions (e.g. epistasis) underlying discrete complex phenotypes (Thompson & Jiggins 2014). Since they were theoretically described in the first half of the XX century, and after the empirical demonstration of their existence in *Drosophila melanogaster* (Darlington & Mather 1949) the understanding of their importance in evolution has increased significantly. Indeed, it has been shown that supergenes underlie highly diverse phenotypic features in a wide range of organisms, such as plant heterostyly in *Primula* (Ernst 1938 in Li et al. 2016), mating type in smut fungi (Bakkeren & Kronstad 1994; Lengeler et al. 2002; Branco et al. 2018) or male sexual morphs in ruffs (Widemo 1998; Küpper et al. 2016) amongst many other examples. These studies highlight the importance of supergenes, not only because they usually underlie ecologically important traits, but also because they can impact our theoretical understanding of evolution. Indeed, the current evolutionary theoretical framework is mostly based on the idea that selection acts on independent loci, and that the vast majority of genetic variance is additive (Orr 2005). Supergenes, however, allow for selection to act on many linked loci at the same time, and for non-additive interactions to be maintained over evolutionary time (Thompson & Jiggins 2014; Schwander et al. 2014). If supergenes play a more important role than previously thought in evolution, the theoretical framework will have to give more room to selection on linked loci and on epistatic traits.

Theory predicts that supergenes would arise whenever selection would favour low recombination rates between two or more loci with epistatic interactions (Turner 1967). That is in situations in which such loci have co-evolved for expressing a given phenotype. In that case, a given locus will be beneficial for the bearer only if it is expressed along with the other co-evolved loci. Recombination and the random spread of such loci will be detrimental, thus favouring low recombination rates between these loci. New mutations arising linked to these loci being beneficial for the bearer would be favoured, forming a supergene which spreads in the population as a single locus (Charlesworth 2016). According to theory, selection for low

recombination is very unlikely to arise between genes that are completely unlinked. In other words, supergenes will arise between genes which are already under some sort of weak linkage, for instance, by being physically close in the genome. It is very unlikely that supergenes could emerge, for example, by translocations that would bring together genes in different chromosomes (Charlesworth & Charlesworth 1975). However, most of these predictions are based on purely theoretical models (e.g. Turner 1967; Charlesworth & Charlesworth 1979; Charlesworth & Charlesworth 1975) and further empirical evidence is needed to critically evaluate these models to fully understand how supergenes form and evolve.

The evolutionary forces favouring selection for low recombination include processes such as the maintenance of local adaptation under strong genetic flow (Tigano & Friesen 2016; Kirkpatrick & Barton 2006) or evolutionary conflict (Charlesworth 2016). In the former case, specific adaptations to a local environment requiring the interactions of two or more alleles would be disrupted by new maladapted variants entering the population through gene flow. Under this scenario, supergenes would maintain the linkage between the locally adapted variants and “shield” them from gene flow. This is the case, for instance, for the different morphs of batesian mimicry in *Heliconius numata* butterflies. In this butterfly species, seven wing colour pattern morphs coexist together in the same populations. These morphs mimic the colour patterns of other butterfly species, which are toxic to predators. *H. numata* thus profits of the signalling for toxicity from other species without the cost of having to produce toxic compounds, a strategy known as Batesian mimicry. Consequently, there are high fitness costs associated with producing the “wrong” wing colour pattern in *H. numata*, as this would result in higher predation rates. Wing colour patterns are controlled by 3 loci, which are co-adapted to produce very specific colour patterns, mimicking different target species (Joron et al. 2006). These three loci are linked together in a large inversion, where recombination rates are extremely low. This arrangement prevents co-adapted alleles within different morphs to recombine in random patterns, even when there is extensive gene flow between morphs (Joron et al. 2011). Another example where gene flow may have played an important role in the emergence of a supergene is the stick insect *Timema cristinae*.This species displays two phenotypes adapted to two different plant species. These differences are heritable and maintained in each ecotype despite gene flow (reviewed in Nosil 2007). One of the key traits involved in ecotype adaptations is the colour and colour patterns of these insects. There is evidence suggesting that these traits are linked to a large inversion in the genome of *T.cristinae*, with low recombination and high genetic differentiation between ecotypes (Lucek et al. 2019). This inversion would potentially keep alleles (in this case,

controlling colour and colour patterns) co-adapted to a specific ecotype to recombine with alleles adapted to a different ecotype in the face of gene flow.

In the case of evolutionary conflict, selection for low recombination would favour the linkage of variants which are beneficial to a specific phenotype and detrimental in another. This is thought to be the case in sex chromosomes, for instance, where male-beneficial alleles are more likely to be linked to the Y chromosome in guppies (Wright et al. 2017). A special case of conflict includes the complexes of selfish genetic elements, such as the Distorter complex in *Drosophila melanogaster*. *D.melanogaster* males carrying alleles with and without the Distorter complex sire almost exclusively offspring carrying the complex. This genetic element is thus an example of gene drive, where a particular allele spreads through the population at rates higher than those expected under Mendelian inheritance (reviewed in Zimmering et al. 1970). The Distorter complex needs the action of two genes to spread, and additionally, 3 other modifier genes are associated with the activity of this drive system. The Distorter complex works as a drive system more efficiently for a specific combination of alleles of these genes. In most Distorter systems, these loci are linked together in inversions that reduce recombination (Thomson & Feldman 1974). If the two main alleles of the complex recombine with non-drive alleles, the Distorter complex is disrupted. This suggests that selection for increased spread of the drive alleles has kept these loci linked together. There is evidence suggesting that the linkage between these alleles has happened several times, with a high rate of turnover in different populations, in a relatively short span of time. This would imply strong selection for linkage in this drive system, even when it is detrimental for the bearer (Brand et al. 2015).

Once selection favours the reduction of recombination, the molecular mechanisms by which it occurs are increasingly becoming understood thanks to the increase in availability and quality of molecular data. These vary from organism to organism and several of them can act at once, they include structural variation such as inversions (e.g. the white-throated sparrow, Thorneycroft 1966; Tuttle et al. 2016) and introgressed regions (e.g. in *Heliconius numata*, Jay et al. 2018), the formation of supergenes in already low recombination areas such as the centromere or epigenetic mechanisms such as methylation (reviewed in Schwander et al. 2014). It is important to note that not all genes linked in a supergene are necessarily involved in defining the phenotype under conflict. Low recombination regions linked to supergenes can expand seemingly without the need of additional selection linkage. This is the case, for example in the supergenes controlling mating type in some fungi. Mating type in fungi requires the joint action of two loci, one emitting a signal and another one receiving it, to enable reproduction and prevent self-fertilisation. These loci are often linked together in a sex chromosome-like system. In the genus *Microbotryum*, this low

recombination region extends in successive inversions, resulting in different evolutionary strata (*i.e.* regions with different levels of differentiation from the homologous chromosome). Intriguingly, only the first inversion involved the linkage of the two mating type loci. The other successive inversions contain no loci associated with mating type, and given that there are no phenotypic differences between mating types, it is unlikely that evolutionary conflict would arise. It is thus likely that the successive inversions and, as a result, the extension of the low recombination region are not driven by adaptive processes, but rather, as a consequence of the lack of recombination (Branco et al. 2017; Branco et al. 2018).

Finally, another important point to consider when it comes to supergenes is the mechanisms by which their polymorphism is maintained. A new variant resulting in a strong phenotypic change would be expected to result in fitness differences. All else being equal, a fitness advantage would lead to the fixation of the new supergene variant and a disadvantage to its loss. Instead, polymorphism can be maintained if the different variants of a supergene are maintained in a population by means of balancing selection (Chouteau et al. 2017) of the phenotypes they encode, by one of the variants being lethal in homozygotes, but where both phenotypes are co-dependent (e.g. sex chromosomes) or a combination of both selection and lethality of homozygotes (Hedrick et al. 2018; Küpper et al. 2016).

The supergene of *Solenopsis invicta*

Solenopsis invicta has a supergene linked to a social polymorphism, where colonies are either single or multiple-queen. This supergene or social chromosome has two variants, SB and Sb between which recombination is suppressed. The queens from multiple-queen colonies are always SB/Sb, whereas queens from single-queen colonies are SB/SB. Sb/Sb results in a lethal recessive, which implies that the Sb variant never recombines. As a consequence of the lack of recombination, the Sb variant shows signs of large-scale gene degenerations such as an increase in non-synonymous mutations (Pracana et al. 2017) and accumulation of repetitive elements (Stolle et al. 2019).

The process by which recombination became suppressed between variants is not yet completely understood. Three large inversions between SB and Sb prevent recombination, but long time gaps between chromosomal rearrangements seem unlikely given that the supergene region has no signs of evolutionary strata (Pracana et al. 2017). This pattern is consistent with the alternative that the Sb variant of the social chromosome introgressed into the *S. invicta* genome from other ant species (Huang et al. 2018). Support for this idea comes from the fact that sister species of *S. invicta* also have the same social chromosome

system. Moreover, the estimated timing of the split between SB and Sb predates the speciation events between many of these species (Wang et al. 2013). In at least some cases the Sb homozygotes are not lethal in these species (Hallar et al. 2007). The introgression hypothesis proposes that the Sb variant was introduced to *S. invicta* as a set of pre-adapted alleles linked together and providing a switch for the multiple-queen colony phenotype (Huang & Wang 2014).

As with other supergenes, an important question is how a stable equilibrium can be maintained between SB and Sb. Homozygote Sb being lethal, all else being equal this variant should be lost in the population. There must therefore be mechanisms at play that prevent this from happening. Multiple hypotheses have been put forward, which are not mutually exclusive. Ecological data show that single and multiple-queen colonies perform better in different environments. Queens from single-queen colonies disperse at longer distances and are able to fund colonies on their own, unlike their multiple-queen counterparts are better at colonising (Ross & Keller 1995a). As a consequence, single-queen colonies are better at becoming established in new, empty environments, whereas multiple-queen colonies are better suited for environments where many other colonies are already established (DeHeer 2002). In the native range of *S. invicta*, seasonal floods are usual, and result in the recurrent disturbance of flood plains, where colonies of *S. invicta* tend to be established (Allen et al. 1974; Buren et al. 1974). This recurrence disturbance provides alternate ecosystems where either single or multiple-queen colonies perform better, thus maintaining a balance between both social forms (Ross & Keller 1995a). Depending on the strength and alternance between these ecological factors, populations of *S. invicta* can contain both types of social forms in relatively balanced proportions, as in the invasive North American range (Callcott & Collins 1996). Alternatively, each social form may show a more patchy distribution, with some populations being almost exclusively formed by one of the colony types, as observed in the native South American populations (Ahrens et al. 2005; Ross et al. 2007; Mescher et al. 2003).

Other hypotheses focus instead on the properties of the social chromosome itself. In multiple-queen colonies most homozygote SB/SB queens are killed by workers before they can leave the nest (DeHeer 2002). Additionally, there is evidence showing that it is mostly workers carrying the Sb allele that attack such queens. This has led to the idea the social chromosome may be an example of theoretical proposal known as the green-beard gene (Keller & Ross 1998). That is, a gene variant (or group of genes) which produces a signal in the carrier that can be detected by other individuals carrying that same variant (Hamilton 1964a; Hamilton 1964b; Dawkins, 1978). In this case, the Sb variant would produce some sort of signal in queens that would be detected by workers, which would in turn recognise

the signal and tolerate the Sb carrying queen. Queens without this variant, on the other hand, would be detected as such by the workers and killed. The idea of Sb as a green-beard element, however, has been contested. Evidence also suggests that workers could be biasing the production of reproductive males towards those carrying the Sb allele (Fritz et al. 2006). The killing of virgin SB/SB queens and the bias against SB carrying males would contribute to maintain the balance between SB and Sb in the population. Finally, there is evidence showing that diploid larvae carrying the Sb allele are more likely to become queens than the SB/SB homozygotes (Buechel et al. 2014). In short, this would imply that individuals carrying the Sb variant has more chances of reproducing than those carrying the SB variant in multiple-queen colonies, thus increasing Sb frequency in the population.

So far, this section has focused on the mechanisms by which recombination is suppressed in the social chromosome, and how its two variants may be maintained at a stable equilibrium in the population. These ideas lead to a key question, forming an important part of this thesis, why does this supergene exist in the first place? Or, put differently, what evolutionary pressures led to the emergence and maintenance of the social chromosome?

The social chromosome emerged from conflict

The relationship between *S. invicta*'s supergene (i.e. social chromosome) and the phenotype to which it is linked is similar to that of sex chromosomes. In chromosomal sex determination systems, there is typically a chromosome which has a large non-recombining region (e.g. Y or W), which occurs in the heterogametic sex (XY males or ZW females), the other sex comprising homogametic individuals (XX females or ZZ males). The homogametic non-recombining chromosome is a lethal recessive (e.g. YY and WW) (Bachtrog et al. 2011). Likewise, in *S. invicta* there are two social chromosomes, one of which has a large region which does not recombine during meiosis (SB and Sb), colony phenotype is determined by whether the queen is homogametic (SB/SB for single-queen colonies) or heterogametic (SB/Sb for multiple-queen colonies) and the non-recombining chromosome is a lethal recessive (Sb/Sb). As with the Y chromosome, the lack of recombination in Sb is associated with a slow process of large-scale gene degeneration (Pracana et al. 2017; Stolle et al. 2019). These similarities between the two systems have led to the idea that maybe the evolutionary forces behind their emergence are the same.

Sexually dimorphic species males and females share most of the genome but have different evolutionary optima. These different 'evolutionary interests' between males and females sometimes result in opposing selection pressures in the genome, a phenomenon referred to as sexual conflict (Chapman et al. 2003). The same process could potentially take place

whenever the same genome expresses two or more different discrete phenotypes within the same population. This broader phenomenon will be referred to as evolutionary conflict throughout this thesis. The emergence of sex chromosomes has often been linked to sexual conflict (Charlesworth 2016). Similarly, a potential evolutionary conflict between the two phenotypes of *S. invicta* could have favoured the emergence of the social chromosome.

For evolutionary conflict to emerge between the two social forms of *S. invicta* selection must act at the colony level, as happens in ants and other eusocial insects. Because all individuals of a colony are highly related, they share most of the alleles in the genome and their evolutionary aims are therefore “aligned” (Hamilton, 1964a; Hamilton, 1964b). As a result, individuals within a colony engage in highly altruistic behaviours that lead to division of labour, where only a few individuals in the colony reproduce. Some authors go as far as to claim that colonies of social insects act as superorganisms, whereby the reproductive caste (e.g. queens) can be viewed as the germline of the colony, and the worker caste the soma (Boomsma & Gawne 2018). Despite this, it is likely that selection also acts at the individual level, within colonies, resulting in conflicts between castes. Such inter-caste conflict, however, is unlikely to explain the emergence of the social chromosome given that all castes are present in both colony types and that the presence of the supergene is not linked to any particular caste. Inter-caste conflict might, however, affect the evolution of the supergene once it has emerged, as detailed later below.

The study of evolutionary conflict in sexually dimorphic species based on entire phenotypes has proved to be extremely challenging, as the relevant traits may be controlled by a wide range of genes and under diverse selection pressures (Mank 2009). Instead, it is easier to focus on single loci under sexual antagonism, a phenomenon known as intralocus sexual conflict (Bonduriansky & Chenoweth 2009). A particular allele at a single locus may be beneficial for one sex but detrimental for the other; in this situation, the phenotype can be shifted towards the optimum (averaged over both sexes) by changing the relative expression of the segregating alleles. In that case, neither of the sexes would be at its optimum. At this point, selection could act to decouple expression levels in the two sexes, allowing sex-biased expression. According to this view, sex-biased expression would represent resolved intralocus sexual conflict (Mank 2009). Sex-biased expression can be achieved by a wide range of mechanisms, including hormonal control (Ketterson et al. 2005). Sex biased expression can be facilitied if the locus is linked to the sexual chromosomes (Bachtrog 2006). Therefore sex chromosomes are expected to be enriched in sex-biased loci (Rice 1984). In organisms which sex is determined by X and Y chromosomes (e.g. *D. melanogaster*), the X chromosome should be enriched in female-biased loci (feminized) and depleted in male-biased loci (de-masculinized). This pattern has been detected in *D.*

melanogaster (Sturgill et al. 2007) and mammals (Khil et al. 2004), and likewise expression on the Z chromosome is de-feminized in birds (Mank & Ellegren 2009), giving support to the idea that the detection of sex-biased expression could act as a tool for detecting intralocus sexual conflict.

However, although informative, this approach has several limitations. Pleiotropic loci (i.e. loci with more than one function) are likely to be under different selective pressures, and would probably not show different expression patterns even if they are under sexual conflict (Mank et al. 2013). Moreover, this approach can only detect resolved sexual conflict, and ongoing sexual conflict would remain undetected (Mank 2009). It is necessary to consider also that sex-biased expression could be neutral for some loci (e.g. a transposon jumping close to a gene could have a cis-acting effect changing the expression of a sex linked locus). But evidence so far supports the expected direction and distribution in the genome of sex biased expression (Khil et al. 2004; Sturgill et al. 2007; Mank & Ellegren 2009). These patterns suggest that considering multiple loci, biased expression will be mostly due to antagonistic selection (Mank 2009). Another issue that needs to be taken into account is dosage compensation. In older sex chromosome systems, the non-recombining variant is often degraded and gene depleted. Therefore, the heterogametic sex has only half of the copy number of the genes compared to the homogametic sex. Consequently, the heterogametic sex has fewer transcription targets, resulting in unbalanced expression levels for all the genes in the sexual chromosome. The expression levels of the two sexes would be evened out by a wide range of mechanisms resulting in dosage compensation (reviewed in Marín et al. 2000). It was once thought that dosage compensation should occur whenever sexual chromosomes evolve. However we now know that dosage compensation can occur at a chromosome-wide level, only locally for some genes, or not at all, thus showing that the evolution of sexual chromosome does not always imply dosage compensation (Mank et al. 2011).

Because social chromosomes share similarities with sex chromosomes, it would be plausible to redirect the techniques that were used to detect markers of intralocus *sexual* conflict, to find intralocus *social* conflict in the social chromosome of *S. invicta*. The SB chromosome would be expected to be enriched in single-queen-biased loci and depleted in multiple-queen-biased loci. Prior to making this prediction, however, it is necessary to consider that the social supergene has a younger age of divergence (>390,000 years according to Wang et al. 2013) than most sex chromosomes studied in which sex-biased selection has been detected (e.g. Y chromosome in humans diverged 80-130 million years ago according to Waters et al. 2001). As a consequence the social chromosomes have had less time to accumulate biased expression than these sexual chromosomes. This shorter time interval

could result in a weaker signal of socially biased expression in the *S. invicta* social chromosomes. The time scale also affects the predictions for dosage compensation: chromosome-level dosage compensation might be less developed in *S. invicta* social chromosomes. Because of the relatively short time of divergence between Sb and SB, the gene content is similar in both variants (Pracana et al. 2017), which would make chromosomal level dosage compensation unnecessary. There are, however, signs of ongoing degeneration in Sb due to lack of recombination, such as the accumulation of potentially deleterious mutations (Pracana et al. 2017) and repetitive elements (Stolle et al. 2019). Consequently, dosage compensation could occur at a gene by gene level. That is, dosage would be compensated only for the few specific genes which are dosage sensitive and for which the Sb variant has been downregulated.

The social chromosome emerged from local adaptation

The previous section discussed the similarities between sex and social chromosomes, and how these could be used to draw parallels between the two systems. There are however, several important differences too between these two supergenes. Differences that could also affect how we interpret the evolutionary forces that originated and maintained the social chromosome.

Sex chromosomes determine the phenotype of discrete individuals whereas social chromosomes determine the phenotype of entire colonies. Ant colonies are neither a simple group of organisms nor a single organism. As explained above, selection in ants acts at the colony level, but it may also act within colony at an individual level. Under the latter scenario there are additional factors and conflicts that may influence the expected outcome of socially biased gene expression and which need to be taken into account. For instance, each caste has different evolutionary aims and therefore the benefits and detriments for either social form will be perceived differently by each caste (Pennell et al. 2018). There are large differences in *S. invicta* gene expression between castes (Ometto et al. 2011), which makes it likely that socially-biased loci will be different between castes. Something similar may occur between developmental stages. Not only is *S. invicta* holometabolous, yielding large phenotypic differences between developmental stages, but also caste is largely defined by social environment during development (Tschinkel 2006). Indeed developmental stage is one of the factors accounting for most of the variation in transcription patterns in ants (Ometto et al. 2011). Similarly studies looking at intralocus sexual conflict at different developmental stages of *Drosophila melanogaster* (which is holometabolous as well) showed that sexual biases in expression occurred at different developmental stages (Perry

et al. 2014). It is thus likely that socially biased loci in *S. invicta* would be found at different developmental stages both due to development and to caste determination.

Additionally, it must be considered that sexual dimorphism does not hinder gene flow between the sexes in any one species, whereas divergence in social form could potentially reduce gene flow (Shoemaker et al. 1996). The studies described above have found that the gene flow between social forms in *S. invicta* is complex, and not necessarily even between phenotypes.

The details of gene flow between *S. invicta* social forms could affect the supergene evolution: for instance, if the queens who populate multiple-queen colonies are predominantly mated by males from single-queen colonies, this asymmetry would produce substantial directional gene flow each generation, from single to multiple-queen colonies (Ross & Shoemaker 1993). It would then be possible that alleles which are favoured in single queen colonies but being slightly deleterious for the multiple-queen phenotype would become fixed population-wide, including the multiple-queen colonies. Put differently, since the entire genome would, on average,have spent more time in single-queen colonies, it would be predominantly adapted to this social phenotype. The only exception to this pattern is the non-recombinant Sb variant of the social chromosome, which is restricted to multiple-queen colonies. In this region of the genome, alleles favoured in multiple-queen colonies would tend to increase in frequency, whatever their effect in single queen colonies. From this perspective linkage between alleles at different loci favoured in multiple-queen alleles could have emerged through a selection scheme similar to models of local adaptation, rather than models of evolutionary conflict. This idea assumes a unidirectional gene flow, from single to multiple-queen colonies, which evidence suggests is inaccurate (Goodisman et al. 2000). Other models of gene flow between social forms are likely to affect the predictions on supergene evolution, and should therefore be considered when attempting to shed light on the processes leading to the emergence and maintenance of the social chromosome.

A final possibility is the emergence of the social chromosome as a selfish supergene. According to the green-beard hypothesis (Keller & Ross 1998), the social chromosome would be acting as a meiotic drive complex. Under this scenario, selection would have pushed to keep the genes for emitting and receiving the “green-beard” signal linked together in the Sb variant. According to this idea, the social chromosome would have first emerged as a selfish genetic element, and its phenotypic effects would have been then hijacked later on (Huang & Wang 2014).

Caste determination in social insects

Termites and many hymenopterans display some degree of social organisation, where individuals within the same colony perform different tasks in a division of labour. In social insects, a caste refers to a group of individuals that performs a similar type of task. The extent to which each individual is specialised in a particular task varies among social insect species but in general these tasks can be roughly divided into reproduction, which will be carried out by one or a few individuals (queens in hymenopterans or queens and kings in termites), and nest maintenance, which will be performed by the bulk of the colony (workers). In some species, castes are flexible, workers retain the ability to reproduce and can become actively reproductive if the queen is removed from the colony (e.g. Chandrashekara & Gadagkar 1991). These are known as behavioural castes, because the differences between workers and queens are mostly behavioural, rather than morphological. In other cases, once established, castes are irreversible and workers are completely unable to reproduce. In these cases castes are different morphologically as well as behaviourally. In social insects with highly complex social organisation, morphological castes reach their highest level of specialisation, to the extent that in some cases castes can be divided into sub-castes, where individuals within the worker caste display morphological differences depending on the specific task they perform (e.g. Sobotník et al. 2010; Gruter et al. 2012; Muscedere & Traniello 2012).

Here, I will be focusing only on ants, where the advent of morphological castes is monophyletic (Hölldobler et al. 1990). Morphological caste differences in ants represent a striking example of complex stable polyphenism. An ant colony can be divided typically into three castes coexisting together in time and space: workers, queens and males. Queens and workers are always female and queens and males share reproductive status, because they are the only individuals that can produce new offspring.

Ants, as well as other hymenopterans, are haplodiploid, which means that males are haploid whereas females (workers and queens) are always diploid. The molecular mechanisms underlying such sex determination system depends on one single locus (Beye et al. 2003; Evans et al. 2004). Individuals that are homozygous or hemizogous for this sex determination locus will develop into males, and heterozygotes become females. There is strong diversifying selection maintaining high genetic diversity for this locus (Hasselmann & Beye 2004). As a consequence, diploid individuals almost always are heterozygote for the sex determining locus and, therefore, devolop into females, whereas haploid individuals,

having only one copy of the gene, will develop into males. In cases of loss of genetic diversity (e.g. a population bottleneck), diploid males become more likely.

The differences between queens and workers are more complex to explore at a molecular level. This is because, with a few exceptions (reviewed in Anderson et al. 2008), caste differences within sex are determined environmentally. More specifically, factors under the control of workers, such as nutritional input during the larval stages will determine whether a female develops into a queen or a worker, or different types of workers (Wheeler 1991; Hölldobler et al. 1990). Ultimately, workers control the proportion of queens and workers in a colony through allocation of resources to larvae. Environmental caste determination implies that all the molecular mechanisms involved in differences between workers and queens are epigenetic, and measurable directly at the gene expression level.

These mechanisms have been the focus of much of the research in social insects. It is only in the last decades, however, with the reduction in price of sequencing, that enough molecular data has been made available for social insects in order to start unraveling the molecular mechanisms underpinning caste differences (Gadau et al. 2012). Initial hypotheses for the molecular machinery behind caste differences borrowed ideas from the field of evolutionary developmental biology. More specifically, the idea that there should be a core of highly connected, highly conserved ‘toolkit’ genes that would work as switches for different caste body plans (Toth & Robinson 2007). Such ‘toolkit’ genes for caste differences have not been found, but there is some empirical support from this idea, stemming from the fact that modules of coexpressed genes do seem to be conserved across ant species (Morandin et al. 2016; Berens et al. 2015). Other authors have suggested that instead, taxonomically restricted young genes with low connectivity are responsible for the morphological innovations involved in caste differences (Sumner 2014). This idea has also gained empirical support (Jasper et al. 2015). The picture that is taking shape is that of a mix of both hypotheses, where both older, highly connected genes and taxonomically restricted genes play a combined role in defining caste differences (Mikheyev & Linksvayer 2015). Under this scenario, the idea that conserved genes have been co-opted to play new roles in the different castes is key. For such genes, rather than changes in gene sequence, molecular changes involved in caste differences would involve changes in the gene regulatory networks (Smith et al. 2015; Simola et al. 2013; Hunt et al. 2011). A potential molecular mechanism that could produce such pattern is that of gene duplication with subfunctionalisation. That is, genes that undergo duplication (locally or as part of larger events), where the new copy of the gene is now under relaxed selection and, therefore, more likely to accumulate new functions (Gadagkar 1997; Chau & Goodisman 2017). Evidence suggests that such mechanisms could be taking place in ants, where several gene

duplications have resulted in expansions of gene families (e.g. Kulmuni et al. 2013; Zhang et al. 2016; Smith et al. 2011), vitellogenins in particular seem to have undergone several duplication events (Morandin et al. 2014; Oxley et al. 2014; Wurm et al. 2011; Corona et al. 2013). In some cases, even potential whole genome duplication events may have taken place during the evolution of some ant groups (Tsutsui et al. 2008). All these duplication events would then provide the potential to respond to selection for subfunctionalisation.

Caste determination in *Solenopsis invicta*

S. invicta, like most ant species, has three morphological castes: queens, males and workers. Sex in fire ants is determined by ploidy, workers and queens are diploid, whereas males are haploid (excepting the case of low genetic diversity outlined earlier). Its worker caste is determined environmentally during the larval stages and they are unable to reproduce (Tschinkel 2006) because of the irreversible degradation of the ovaries. In *S. invicta* there are no discrete phenotypic differences within the worker caste. There is, however, a continuum in size variation (Tschinkel 1988; Tschinkel 2006). Despite the absence of discrete morphological subcastes, there is a high level of task specialisation based on the size of the workers. Smaller workers tend to perform tasks inside the nest, like brood maintenance or tunnel building, whereas larger workers tend to perform tasks outside the nest, including defense and foraging (Porter & Tschinkel 1985).

Even though all castes share the same genome, the expression patterns between them are very different. Overall, males and workers show the most expression differences, and males and queens the least. This pattern might have been expected, as males and queens share reproductive status, workers and queens share sex, but workers and males share neither. Additionally, workers show the most genes with higher expression, a pattern that has been interpreted as a consequence of the different environments that workers have to cope with (Ometto et al. 2011). Within workers, specific behaviours (e.g. nursing vs foraging) are also detectable through specific expression patterns. And these differences are known to disappear in the absence of a queen (Manfredini et al. 2014).

No single gene has been found to determine caste in any ant, and *S. invicta* is not an exception. There are, however specific genes and gene families that are promising candidates to play a key role in caste differences. The ground-plan hypothesis (West-Eberhard 1987) establishes that cases could have evolved by decoupling reproductive behaviours in workers (feeding and nest maintenance) and queens (egg production). This hypothesis has gained considerable empirical support in other social insects, including ants

(e.g. Kapheim & Johnson 2017; Ihle et al. 2010; Roy-Zokan et al. 2015; Pamminger & Hughes 2017). At the molecular level, under this hypothesis we would expect that genes involved in reproduction in other species would play an important role in caste differences. Additionally, we would expect many of these genes to have been duplicated, allowing for caste subfunctionalisation. Supporting this idea, genes involved in sex determination in *Drosophila melanogaster* such as double-sex or fruitless have been shown to display caste-specific splice forms in *S. invicta*. Moreover, the fire ant has 4 copies of the vitellogenin gene, which is usually involved in yolk production for eggs in other arthropods. These copies also display caste-specific expression patterns (Wurm et al. 2011).

In ants and other social insects, most of the communication between individuals that keeps the social structure of the colony is done through pheromones (Hölldobler et al. 1990). If chemical communication is disrupted, colonies collapse (Yan et al. 2017; Trible et al. 2017). Genes involved in chemical communication such as those encoding for odorant binding proteins (Zhang et al. 2016; Pracana et al. 2017) or chemosensory proteins (Koch et al. 2013; Hojo et al. 2015) are therefore likely to be relevant for caste differences.

Even though the molecular basis of caste differences are becoming clearer, there is a danger that these results include experimental artifacts. This is because most expression pattern studies carried out in *S. invicta* have used whole bodies, which can result in misleading expression patterns arising from changes in allometry between castes (Johnson et al. 2013), or just a single or a handful of tissues. As a result, it is not possible to fully understand the molecular mechanisms underlying caste differences with the available data. High resolution tissue-specific gene expression data is needed in order to account for allometric changes between castes, while unraveling the caste-specific functions of separate tissues.

Aims and objectives

The red fire ant *Solenopsis invicta* displays two types of social organisation. This social polymorphism is controlled by a single genetic element, the “social chromosome”. A large part of this social chromosome is a supergene, a region of the genome with repressed recombination that links together hundreds of genes, which spread through the population as a single Mendelian element. Additionally, because ants produce three different castes from the same genome, they represent a good example of polyphenism, whereby the same genotype produces different phenotypes. The red fire ant is, therefore, an exceptional system for asking questions about the interplay between genomes and phenotypes. My project combines analyses of gene expression patterns and genetic modelling to answer some of these questions. More specifically:

In **chapter 2** I use expression patterns between variants of the supergene and between social forms to unravel the evolutionary forces that may have originated and maintained the evolution of the social chromosome.
In **chapter 3** I model analytically the spread of an allele with different fitness effects between social forms. Different gene flows are modelled, with and without linkage to the supergene to test under which conditions evolution by conflict or by local adaptation is favoured.
In **chapter 4** I use tissue-specific gene expression patterns between 16 and 19 tissues of males, workers and queens of *S. invicta* to explore the molecular underpinning of caste differences.

References

Ahrens, M.E., Ross, K.G. & Shoemaker, D.D., 2005. Phylogeographic structure of the fire ant Solenopsis invicta in its native South American range: roles of natural barriers and habitat connectivity. *Evolution; international journal of organic evolution*, 59(8), p.1733–1743.

Allen, G.E. et al., 1974. The red imported fire ant, *Solenopsis invicta*; Distribution and habitat in Mato Grosso, Brazil. *Annals of the Entomological Society of America*, 67(1), p.43–46.

Anderson, K.E., Linksvayer, T.A. & Smith, C.R., 2008. The causes and consequences of genetic caste determination in ants (Hymenoptera: Formicidae). *Myrmecological news / Osterreichische Gesellschaft fur Entomofaunistik*, 11, p.119–132.

Ascunce, M.S. et al., 2011. Global invasion history of the fire ant *Solenopsis invicta*. *Science*, 331(6020), p.1066–1068.

Ayali, A., 2019. The puzzle of locust density-dependent phase polyphenism. *Current opinion in insect science*, 35, p.41–47.

Bachtrog, D., 2006. A dynamic view of sex chromosome evolution. *Current opinion in genetics & development*, 16(6), p.578–585.

Bachtrog, D. et al., 2011. Are all sex chromosomes created equal? *Trends in Genetics*, 27(9), p.350–357.

Bakkeren, G. & Kronstad, J.W., 1994. Linkage of mating-type loci distinguishes bipolar from tetrapolar mating in basidiomycetous smut fungi. *Proceedings of the National Academy of Sciences of the United States of America*, 91(15), p.7085–7089.

Berens, A.J., Hunt, J.H. & Toth, A.L., 2015. Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. *Molecular biology and evolution*, 32(3), p.690–703.

Beye, M. et al., 2003. The gene csd is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell*, 114(4), p.419–429.

Bonduriansky, R. & Chenoweth, S.F., 2009. Intralocus sexual conflict. *Trends in ecology & evolution*, 24(5), p.280–288.

Boomsma, J.J. & Gawne, R., 2018. Superorganismality and caste differentiation as points of

no return: how the major evolutionary transitions were lost in translation. *Biological reviews of the Cambridge Philosophical Society*, 93(1), p.28–54.

Branco, S. et al., 2017. Evolutionary strata on young mating-type chromosomes despite the lack of sexual antagonism. *Proceedings of the National Academy of Sciences of the United States of America*, 114(27), p.7067–7072.

Branco, S. et al., 2018. Multiple convergent supergene evolution events in mating-type chromosomes. *Nature Communications*, 9(1)

Brand, C.L., Larracuente, A.M. & Presgraves, D.C., 2015. Origin, evolution, and population genetics of the selfish Segregation Distorter gene duplication in European and African populations of *Drosophila melanogaster*. *Evolution; international journal of organic evolution*, 69(5), p.1271–1283.

Buechel, S.D., Wurm, Y. & Keller, L., 2014. Social chromosome variants differentially affect queen determination and the survival of workers in the fire ant *Solenopsis invicta*. *Molecular ecology*, 23(20), p.5117–5127.

Buren, W.F. et al., 1974. Zoogeography of the Imported Fire Ants. *Journal of the New York Entomological Society*, 82(2), p.113–124.

Callcott, A.-M.A. & Collins, H.L., 1996. Invasion and Range Expansion of Imported Fire Ants (Hymenoptera: Formicidae) in North America from 1918-1995. *The Florida entomologist*, 79(2), p.240–251.

Chandrashekara, K. & Gadagkar, R., 1991. Behavioural castes, dominance and division of labour in a primitively eusocial wasp. *Ethology: formerly Zeitschrift fur Tierpsychologie*, 87(3-4), p.269–283.

Chapman, T. et al., 2003. Sexual conflict. *Trends in ecology & evolution*, 18(1), p.41–47.

Charlesworth, D., 2016. The status of supergenes in the 21st century: recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. *Evolutionary applications*, 9(1), p.74–90.

Charlesworth, D. & Charlesworth, B., 1979. Selection on recombination in a multi-locus system. *Genetics*, 91(3), p.575–580.

Charlesworth, D. & Charlesworth, B., 1975. Theoretical genetics of Batesian mimicry II. Evolution of supergenes. *Journal of theoretical biology*, 55(2), p.305–324.

Chau, L.M. & Goodisman, M.A.D., 2017. Gene duplication and the evolution of phenotypic diversity in insect societies. *Evolution; international journal of organic evolution*, 71(12), p.2871–2884.

Chirino, M.G., Gilbert, L.E. & Folgarait, P.J., 2012. Behavioral discrimination between monogyne and polygyne red fire ants (Hymenoptera: Formicidae) in Their Native Range. *Annals of the Entomological Society of America*, 105(5), p.740–745.

Chouteau, M. et al., 2017. Polymorphism at a mimicry supergene maintained by opposing frequency-dependent selection pressures. *Proceedings of the National Academy of Sciences of the United States of America*, 114(31), p.8325–8329.

Corona, M. et al., 2013. Vitellogenin underwent subfunctionalization to acquire caste and behavioral specific expression in the harvester ant *Pogonomyrmex barbatus*. *PLoS genetics*, 9(8), p.1003730.

Darlington, C.D. & Mather, K., 1949. *The elements of genetics*, George Allen & Unwin Ltd: London.

DeHeer, C.J., 2002. A comparison of the colony-founding potential of queens from single- and multiple-queen colonies of the fire ant *Solenopsis invicta*. *Animal behaviour*, 64(4), p.655–661.

Ernst, A., 1938. Weitere Untersuchungen zur Phänanalyse, zum Fertilitätsproblem und zur Genetik heterostyler Primeln: Die *F_1tn1-Bastarde Pr.(hortensis viscosa)*, Art. Institut Orell Füssli [Abt. Zeitschriften].

Evans, J.D., Shearman, D.C.A. & Oldroyd, B.P., 2004. Molecular basis of sex determination in haplodiploids. *Trends in ecology & evolution*, 19(1), p.1–3.

Fritz, G.N., Vander Meer, R.K. & Preston, C.A., 2006. Selective male mortality in the red imported fire ant, *Solenopsis invicta*. *Genetics*, 173(1), p.207–213.

Gadagkar, R., 1997. The evolution of caste polymorphism in social insects: Genetic release followed by diversifying evolution. *Journal of genetics*, 76(3), p.167–179.

Gadau, J. et al., 2012. The genomic impact of 100 million years of social evolution in seven ant species. *Trends in genetics: TIG*, 28(1), p.14–21.

Goodisman, M.A., Ross, K.G. & Asmussen, M.A., 2000. A formal assessment of gene flow and selection in the fire ant *Solenopsis invicta*. *Evolution; international journal of organic*

evolution, 54(2), p.606–616.

Gotzek, D. & Ross, K.G., 2007. Genetic regulation of colony social organization in fire ants: an integrative overview. *The Quarterly review of biology*, 82(3), p.201–226.

Greenberg, L., Fletcher, D.J.C. & Vinson, S.B., 1985. Differences in worker size and mound distribution in monogynous and polygynous colonies of the fire ant *Solenopsis invicta* Buren. *Journal of the Kansas Entomological Society*, 58(1), p.9–18.

Gruter, C. et al., 2012. A morphologically specialized soldier caste improves colony defense in a neotropical eusocial bee. *Proceedings of the National Academy of Sciences*, 109(4), p.1182–1186.

Hallar, B.L., Krieger, M.J.B. & Ross, K.G., 2007. Potential cause of lethality of an allele implicated in social evolution in fire ants. *Genetica*, 131(1), p.69–79.

Hamilton, W.D., 1964a. The genetical evolution of social behaviour. I. *Journal of theoretical biology*, 7(1), p.1–16.

Hamilton, W.D., 1964b. The genetical evolution of social behaviour. II. *Journal of theoretical biology*, 7(1), p.17–52.

Hasselmann, M. & Beye, M., 2004. Signatures of selection among sex-determining alleles of the honey bee. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14), p.4888–4893.

Hedrick, P.W., Tuttle, E.M. & Gonser, R.A., 2018. Negative-assortative mating in the white-throated sparrow. *Journal of Heredity*, 109(3), p.223–231.

Hojo, M.K. et al., 2015. Antennal RNA-sequencing analysis reveals evolutionary aspects of chemosensory proteins in the carpenter ant, *Camponotus japonicus*. *Scientific reports*, 5, p.13541.

Hölldobler, B. et al., 1990. *The Ants*, Harvard University Press.

Huang, Y.-C. et al., 2018. Multiple large inversions and breakpoint rewiring of gene expression in the evolution of the fire ant social supergene. *Proceedings of the Royal Society B: Biological Sciences*, 285(1878).

Huang, Y.-C. & Wang, J., 2014. Did the fire ant supergene evolve selfishly or socially? *BioEssays: news and reviews in molecular, cellular and developmental biology*, 36(2), p.200–208.

Hunt, B.G. et al., 2011. Relaxed selection is a precursor to the evolution of phenotypic plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, 108(38), p.15936–15941.

Ihle, K.E. et al., 2010. Genotype effect on regulation of behaviour by vitellogenin supports reproductive origin of honeybee foraging bias. *Animal behaviour*, 79(5), p.1001–1006.

Jasper, W.C. et al., 2015. Large-scale coding sequence change underlies the evolution of postdevelopmental novelty in honey bees. *Molecular biology and evolution*, 32(2), p.334–346.

Jay, P. et al., 2018. Supergene evolution triggered by the introgression of a chromosomal inversion. *Current biology: CB*, 28(11), p.1839–1845.e3.

Johnson, B.R., Atallah, J. & Plachetzki, D.C., 2013. The importance of tissue specificity for RNA-seq: highlighting the errors of composite structure extractions. *BMC genomics*, 14, p.586.

Joron, M. et al., 2006. A conserved supergene locus controls colour pattern diversity in *Heliconius* butterflies. *PLoS biology*, 4(10), p.303.

Joron, M. et al., 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363), p.203–206.

Kang, L. et al., 2004. The analysis of large-scale gene expression correlated to the phase changes of the migratory locust. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51), p.17611–17615.

Kapheim, K.M. & Johnson, M.M., 2017. Support for the reproductive ground plan hypothesis in a solitary bee: links between sucrose response and reproductive status. *Proceedings of the Royal Society B: Biological Sciences*, 284(1847).

Keller, L., 1995. Social life: the paradox of multiple-queen colonies. *Trends in ecology & evolution*, 10(9), p.355–360.

Keller, L. & Ross, K.G., 1998. Selfish genes: a green beard in the red fire ant. *Nature*, 394, p.573.

Ketterson, E.D., Nolan, V., Jr & Sandell, M., 2005. Testosterone in females: mediator of adaptive traits, constraint on sexual dimorphism, or both? *The American naturalist*, 166 Suppl 4, p.S85–98.

Khil, P.P. et al., 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nature genetics*, 36(6), p.642–646.

Kirkpatrick, M. & Barton, N., 2006. Chromosome inversions, local adaptation and speciation. *Genetics*, 173(1), p.419–434.

Koch, S.I. et al., 2013. Caste-specific expression patterns of immune response and chemosensory related genes in the leaf-cutting ant, *Atta vollenweideri*. *PloS one*, 8(11), p.81518.

Krieger, M.J.B. & Ross, K.G., 2002. Identification of a major gene regulating complex social behavior. *Science*, 295(5553), p.328–332.

Kulmuni, J., Wurm, Y. & Pamilo, P., 2013. Comparative genomics of chemosensory protein genes reveals rapid evolution and positive selection in ant-specific duplicates. *Heredity*, 110(6), p.538–547.

Küpper, C. et al., 2016. A supergene determines highly divergent male reproductive morphs in the ruff. *Nature genetics*, 48(1), p.79–83.

Lackey, A.C.R. et al., 2019. Lifetime fitness, sex-specific life history, and the maintenance of a polyphenism. *The American naturalist*, 194(2), p.230–245.

Lawson, L.P., Vander Meer, R.K. & Shoemaker, D., 2012. Male reproductive fitness and queen polyandry are linked to variation in the supergene Gp-9 in the fire ant *Solenopsis invicta*. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741), p.3217–3222.

Lengeler, K.B. et al., 2002. Mating-type locus of Cryptococcus neoformans: a step in the evolution of sex chromosomes. *Eukaryotic cell*, 1(5), p.704–718.

Li, J. et al., 2016. Genetic architecture and evolution of the S locus supergene in *Primula vulgaris*. *Nature plants*, 2(12), p.16188.

Lucek, K., Gompert, Z. & Nosil, P., 2019. The role of structural genomic variants in population differentiation and ecotype formation in *Timema cristinae* walking sticks. *Molecular ecology*, 28(6), p.1224–1237.

Manfredini, F. et al., 2014. Molecular and social regulation of worker division of labour in fire ants. *Molecular ecology*, 23(3), p.660–672.

Mank, J.E., 2009. Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. *The American naturalist*, 173(2), p.141–150.

Mank, J.E. & Ellegren, H., 2009. Sex-linkage of sexually antagonistic genes is predicted by female, but not male, effects in birds. *Evolution; international journal of organic evolution*, 63(6), p.1464–1472.

Mank, J.E., Hosken, D.J. & Wedell, N., 2011. Some inconvenient truths about sex chromosome dosage compensation and the potential role of sexual conflict. *Evolution; international journal of organic evolution*, 65(8), p.2133–2144.

Mank, J.E., Wedell, N. & Hosken, D.J., 2013. Polyandry and sex-specific gene expression. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1613), p.20120047.

Marín, I., Siegal, M.L. & Baker, B.S., 2000. The evolution of dosage-compensation mechanisms. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 22(12), p.1106–1114.

Matsumoto, K. & Knell, R.J., 2017. Diverse and complex male polymorphisms in *Odontolabis* stag beetles (Coleoptera: *Lucanidae*). *Scientific reports*, 7(1), p.16733.

Mayr, E., 1982. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*, Harvard University Press.

Mescher, M.C. et al., 2003. Distribution of the two social forms of the fire ant *Solenopsis invicta* (Hymenoptera: Formicidae) in the native South American range. *Annals of the Entomological Society of America*, 96(6), p.810–817.

Mikheyev, A.S. & Linksvayer, T.A., 2015. Genes associated with ant social behavior show distinct transcriptional and evolutionary patterns. *eLife*, 4, p.04775.

Morandin, C. et al., 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome biology*, 17, p.43.

Morandin, C. et al., 2014. Not only for egg yolk—functional and evolutionary insights from expression, selection, and structural analyses of Formica ant vitellogenins. *Molecular biology and evolution*, 31(8), p.2181–2193.

Muscedere, M.L. & Traniello, J.F.A., 2012. Division of labor in the hyperdiverse ant genus

Pheidole is associated with distinct subcaste- and age-related patterns of worker brain organization. *PLoS ONE*, 7(2), p.31618.

Nevo, E., 2001. Evolution of genome–phenome diversity under environmental stress. *Proceedings of the National Academy of Sciences of the United States of America*, 98(11), p.6233–6240.

Nosil, P., 2007. Divergent host plant adaptation and reproductive isolation between ecotypes of *Timema cristinae* walking sticks. *The American naturalist*, 169(2), p.151–162.

Ometto, L. et al., 2011. Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. *Molecular biology and evolution*, 28(4), p.1381–1392.

Orr, H.A., 2005. The genetic theory of adaptation: a brief history. *Nature reviews. Genetics*, 6(2), p.119–127.

Oxley, P.R. et al., 2014. The genome of the clonal raider ant *Cerapachys biroi*. *Current biology: CB*, 24(4), p.451–458.

Pamminger, T. & Hughes, W.O.H., 2017. Testing the reproductive ground plan hypothesis in ants (Hymenoptera: Formicidae). *Evolution; international journal of organic evolution*, 71(1), p.153–159.

Pennell, T.M. et al., 2018. Building a new research framework for social evolution: intralocus caste antagonism. *Biological reviews of the Cambridge Philosophical Society*, 93(2), p.1251–1268.

Perry, J.C., Harrison, P.W. & Mank, J.E., 2014. The ontogeny and evolution of sex-biased gene expression in *Drosophila melanogaster*. *Molecular biology and evolution*, 31(5), p.1206–1219.

Porter, S.D. & Tschinkel, W.R., 1985. Fire ant polymorphism: the ergonomics of brood production. *Behavioral ecology and sociobiology*, 16(4), p.323–336.

Pracana, R. et al., 2017. The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. *Molecular ecology*, 26(11), p.2864–2879.

Rice, W.R., 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution; international journal of organic evolution*, 38(4), p.735–742.

Ross, K.G. et al., 1993. Effect of a founder event on variation in the genetic sex-determining

system of the fire ant *Solenopsis invicta*. *Genetics*, 135(3), p.843–854.

Ross, K.G. et al., 2007. Genetic variation and structure in native populations of the fire ant *Solenopsis invicta*: evolutionary and demographic implications. *Biological journal of the Linnean Society. Linnean Society of London*, 92(3), p.541–560.

Ross, K.G. & Fletcher, D.J.C., 1986. Diploid male production — a significant colony mortality factor in the fire ant *Solenopsis invicta* (Hymenoptera: Formicidae). *Behavioral ecology and sociobiology*, 19(4), p.283–291.

Ross, K.G. & Keller, L., 1995a. Ecology and evolution of social organization: insights from fire ants and other highly eusocial insects. *Annual review of ecology and systematics*, 26(1), p.631–656.

Ross, K.G. & Keller, L., 1995b. Joint influence of gene flow and selection on a reproductively important genetic polymorphism in the fire ant *Solenopsis invicta*. *The American naturalist*, 146(3), p.325–348.

Ross, K.G. & Shoemaker, D.D., 1993. An unusual pattern of gene flow between the two social forms of the fire ant *Solenopsis invicta*. *Evolution; international journal of organic evolution*, 47(5), p.1595–1605.

Ross, K.G. & Shoemaker, D.D., 1997. Nuclear and mitochondrial genetic structure in two social forms of the fire ant *Solenopsis invicta*: insights into transitions to an alternate social organization. *Heredity*, 78, p.590.

Ross, K. & Keller, L., 2002. Experimental conversion of colony social organization by manipulation of worker genotype composition in fire ants (*Solenopsis invicta*). *Behavioral ecology and sociobiology*, 51(3), p.287–295.

Roy-Zokan, E.M. et al., 2015. Vitellogenin and vitellogenin receptor gene expression is associated with male and female parenting in a subsocial insect. *Proceedings of the Royal Society B: Biological Sciences*, 282(1809), p.20150787.

Saddoris, K., Fritz, A.H. & Fritz, G.N., 2016. Evidence of selective mating and triploidy among two social forms of *Solenopsis invicta* (Hymenoptera: Formicidae). *The Florida entomologist*, 99(3), p.566–568.

Schwander, T., Libbrecht, R. & Keller, L., 2014. Supergenes and complex phenotypes. *Current biology: CB*, 24(7), p.R288–94.

Shoemaker, D.D., DeWayne Shoemaker, D. & Ross, K.G., 1996. Effects of social organization on gene flow in the fire ant *Solenopsis invicta*. *Nature*, 383(6601), p.613–616.

Simola, D.F. et al., 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome research*, 23(8), p.1235–1247.

Smith, C.D. et al., 2011. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), p.5673–5678.

Smith, C.R. et al., 2015. How do genomes create novel phenotypes? Insights from the loss of the worker caste in ant social parasites. *Molecular biology and evolution*, 32(11), p.2919–2931.

Sobotník, J., Jirosová, A. & Hanus, R., 2010. Chemical warfare in termites. *Journal of insect physiology*, 56(9), p.1012–1021.

Stolle, E. et al., 2019. Degenerative expansion of a young supergene. *Molecular biology and evolution*, 36(3), p.553–561.

Sturgill, D. et al., 2007. Demasculinization of X chromosomes in the *Drosophila* genus. *Nature*, 450(7167), p.238–241.

Sumner, S., 2014. The importance of genomic novelty in social evolution. *Molecular ecology*, 23(1), p.26–28.

Thompson, M.J. & Jiggins, C.D., 2014. Supergenes and their role in evolution. *Heredity*, 113(1), p.1–8.

Thomson, G.J. & Feldman, M.W., 1974. Population genetics of modifiers of meiotic drive. II. Linkage modification in the segregation distortion system. *Theoretical population biology*, 5(2), p.155–162.

Thorneycroft, H.B., 1966. Chromosomal polymorphism in the white-throated sparrow, *Zonotrichia albicollis* (Gmelin). *Science*, 154(3756), p.1571–1572.

Tigano, A. & Friesen, V.L., 2016. Genomics of local adaptation with gene flow. *Molecular ecology*, 25(10), p.2144–2164.

Toth, A.L. & Robinson, G.E., 2007. Evo-devo and the evolution of social behavior. *Trends in*

genetics: TIG, 23(7), p.334–341.

Trible, W. et al., 2017. orco mutagenesis causes loss of antennal lobe glomeruli and impaired social behavior in ants. *Cell*, 170(4), p.727–735.e10.

Tschinkel, W.R., 1988. Colony growth and the ontogeny of worker polymorphism in the fire ant, Solenopsis invicta. *Behavioral ecology and sociobiology*, 22(2), p.103–115.

Tschinkel, W.R., 2006. *The Fire Ants*, Harvard University Press.

Tsutsui, N.D. et al., 2008. The evolution of genome size in ants. *BMC evolutionary biology*, 8, p.64.

Turner, J.R.G., 1967. On supergenes. I. The evolution of supergenes. *The American naturalist*, 101(919), p.195–221.

Tuttle, E.M. et al., 2016. Divergence and Functional Degradation of a Sex Chromosome-like Supergene. *Current biology: CB*, 26(3), p.344–350.

Vargo, E.L. & Porter, S.D., 1989. Colony reproduction by budding in the polygyne form of *Solenopsis invicta* (Hymenoptera: Formicidae). *Annals of the Entomological Society of America*, 82(3), p.307–313.

Wang, J. et al., 2013. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature*, 493(7434), p.664–668.

Waters, P.D. et al., 2001. The human Y chromosome derives largely from a single autosomal region added to the sex chromosomes 80–130 million years ago. *Cytogenetic and Genome Research*, 92(1-2), p.74–79. .

West-Eberhard, M.J., 1987. Flexible strategy and social evolution. *Animal societies. Theories and facts.*.

Wheeler, D.E., 1991. The Developmental Basis of Worker Caste Polymorphism in Ants. *The American naturalist*, 138(5), p.1218–1238.

Widemo, F., 1998. Alternative reproductive strategies in the ruff, *Philomachus pugnax*: a mixed ESS? *Animal behaviour*, 56(2), p.329–336.

Wright, A.E. et al., 2017. Convergent recombination suppression suggests role of sexual selection in guppy sex chromosome formation. *Nature communications*, 8, p.14251.

Wurm, Y. et al., 2011. The genome of the fire ant *Solenopsis invicta. Proceedings of the National Academy of Sciences of the United States of America*, 108(14), p.5679–5684.

Yan, H. et al., 2017. An Engineered orco mutation produces aberrant social behavior and defective neural development in ants. *Cell*, 170(4), p.736–747.e9.

Zhang, W. et al., 2016. Tissue, developmental, and caste-specific expression of odorant binding proteins in a eusocial insect, the red imported fire ant, *Solenopsis invicta*. *Scientific reports*, 6, p.35452.

Zimmering, S., Sandler, L. & Nicoletti, B., 1970. Mechanisms of meiotic drive. *Annual review of genetics*, 4, p.409–436.

Chapter 2: Genomic architecture and evolutionary conflict drive allele-specific expression in a social supergene

Collaborations in this chapter

Dr. Eckart Stolle collected the fire ant samples from the South American populations and generated the DNAseq data for the samples used in variant calling

Dr. Rodrigo Pracana performed the variant calling based on the DNAseq data and generated the VCF files on which the allele-specific expression analyses are based.

Dr. Monika Struebig generated the RNAseq libraries for the South American populations

I performed the rest of lab work and analyses described here and wrote this chapter.

Abstract

Supergene regions - where recombination is suppressed over large parts of the genome which can contain hundreds of genes - have recently been shown to be pervasive in defining and maintaining complex polymorphisms in a wide range of organisms. Here, we use the young “social chromosome” supergene system of the red fire ant *Solenopsis invicta* to test the role of intra-genomic conflict in young supergene evolution. This supergene has two variants, SB and Sb that determine a trait linked to alternative reproductive strategies: the number of queens per colony. We hypothesise that the forces shaping the evolution of this supergene are a consequence of evolutionary conflict, where each phenotype has a different optimum yet share most of the genome. Evolutionary conflict would impact gene expression within the supergene resulting in particular patterns of expression.

To test whether this is the case, we perform extensive comparisons of gene expression between single- and multiple-queen colonies, and in heterozygous individuals found only in multiple-queen colonies, between SB and Sb alleles of the hundreds of genes of the supergene region.

We find that the majority of gene expression differences between supergene variants are driven by the genomic architecture of the social chromosome. In contrast, a much smaller portion of gene expression differences have the potential to be adaptive and due to evolutionary conflict between social forms. Overall, our results demonstrate that supergene evolution is a complex process due to a combination of neutral, deleterious and adaptive mechanisms.

Introduction

Supergenes are paradigmatic examples of how genome structure can drastically impact gene interactions to produce large phenotypic differences within species. At an extreme level, supergenes are low recombination regions of the genome linked to the expression of discrete, complex polymorphisms within species. They contain tightly linked alleles for up to hundreds of genes, allowing for the maintenance of genetic interactions over evolutionary time (Darlington & Mather 1949). These regions are believed to emerge when several phenotypes with different evolutionary optima are connected by gene flow. The ensuing evolutionary conflict would result in selection pressures for low recombination of co-adapted alleles which together encode for the different phenotypes (Charlesworth 2016). Supergenes are more widespread than previously thought and they control ecologically important traits in a wide variety of organisms, including flower heterostyly in *Primula* (Li et al. 2016), mating type in *Mycrobotryum* fungi (Branco et al. 2018) or male sexual morphs in ruffs (Küpper et al. 2016).

The non-recombining portions of sex chromosomes represent a special type of supergene, because the two sexes are interdependent (Bergero & Charlesworth 2009). There is a large body of research on the evolution of sex chromosomes, which constitutes a good starting point to understand the evolution of supergenes in general. In these systems, evolutionary conflict over sexual phenotypes (sexual conflict) is believed to cause an enrichment of sexually-biased loci in the sex chromosomes, where the Y (or Z) is masculinized, and the X (or W) variant feminized (Wright et al. 2017; Mank 2017), a pattern observed in a wide range of cases (Zemp et al. 2016; Vicoso et al. 2013; Khil et al. 2004; Parsch & Ellegren 2013). On the other hand, studies focusing on systems with young sex chromosomes or where there is a high rate of sex chromosomes turnover (e.g. Dufresnes et al. 2015; Nozawa et al. 2014; Alekseyenko et al. 2013; Muyle et al. 2012) suggest that the role of evolutionary conflict in driving sex chromosome evolution might have been overestimated. Instead, sex chromosome evolution could be driven mostly by processes unrelated to defining phenotypic sex differences (Cavoto et al. 2018; Branco et al. 2018; Branco et al. 2017). Such processes include gene-specific dosage compensation due to the accumulation of deleterious mutations in the non-recombining variant (Bachtrog 2013; Gu & Walters 2017), or the random insertion of transposable elements (Stolle et al. 2019; Ellison & Bachtrog 2018), that can in turn affect expression patterns (Cowley & Oakey 2013). In sum, the effects of selection for alternate genomic optima to overcome intra-genomic conflict and the effects that are phenotypically neutral are relatively important during the different stages of sex chromosome evolution (Bachtrog 2013).

To test the relative importance of evolutionary conflict and processes related to recombination suppression shaping the early evolution of an ecologically important supergene, we focus on the social supergene of the red fire ant *Solenopsis invicta*. Two social forms coexist in this species: colonies either have one or multiple queens. This social polymorphism is associated with multiple behavioral and physiological traits, all of which are controlled by two variants of a supergene known as 'social chromosome' (Wang et al. 2013). Recombination is severely repressed between the two variants of the social chromosome, SB and Sb, due to the presence of at least 3 inversions (Stolle et al. 2019). All individuals in single queen colonies are homozygotes SB/SB. In multiple queen colonies, all egg-laying queens are SB/Sb but workers can either be homozygotes or heterozygotes. Sb/Sb individuals are rarely found in the wild and are believed to be lethal recessives (Gotzek & Ross 2007; Fritz et al. 2006), resulting in suppressed recombination in Sb. The two variants diverged less than a million years ago (Stolle et al. 2019). In line with such a short time of divergence, the social chromosome shows genomic patterns that evidence the ongoing degeneration of Sb, by accumulating potentially deleterious mutations due to the inefficient selection caused by the lack of recombination (Stolle et al. 2019; Pracana et al. 2017). However the vast majority of genes are found intact in both social chromosomes - likely because of a potential high cost of loss of genes during the early evolution of this system (Stolle et al. 2019).

We tested whether expression patterns for genes in the *S. invicta* supergene region follow what would be predicted under a scenario led by conflict or by phenotypically-neutral evolutionary processes. For this, we generated a new RNAseq dataset from SB/Sb individuals of populations in the native range of *S. invicta* in South America, using different body parts of queens (head, thorax and abdomen) and whole workers. We additionally integrate preexisting RNAseq data generated from whole bodies of queens from the invasive range of *S. invicta* in North America in our analysis.

We identified genes with fixed differences between the SB and Sb variants of the social chromosome across different populations of *S. invicta*. We then used this information to detect allele-specific expression differences between variants, and compared these expression patterns with those obtained from comparisons between social forms. Our results show that both evolutionary conflict and phenotypically-neutral processes are likely to play a role in the evolution of young supergenes. Their relative importance in gene expression evolution differs, however, with phenotypically-neutral processes explaining most of the expression patterns observed.

Methods

Generation of RNAseq gene expression data

We used two previously published and generated one novel *S. invicta* RNAseq datasets. Wurm et al. (2011) obtained whole-body RNAseq data from six pools of 4 egg-laying SB/Sb queens, each from a multiple-queen colony from Georgia, USA. Morandin et al. (2016) obtained whole-body RNAseq data from 6 samples, each being a pool of 3 queens from single- or a multiple-queen colonies (3 replicates per social form) from Texas, USA. Due to low quality, we eliminated one queen sample from a multiple-queen colony. All the queens were mature and egg-laying (Morandin personal communication), thus queens from multiple queen colonies carried the SB/Sb genotype (Keller and Ross 1998). More details for both datasets can be found in Annex I (Table AI.1a).

Both published datasets are from pools of whole bodies from the invasive North American range of *S. invicta.* Because comparisons of whole bodies can be confounded by allometric differences (Johnson et al. 2013) and genetic diversity is reduced among Sb haplotypes in the invasive populations of North America (Pracana, Priyam, et al. 2017), we generated a new gene expression dataset. From the native Argentinian range of this species, we collected 6 multiple-queen colonies in 2014 (collection and exportation permit numbers 007/15, 282/2016, 433/02101-0014449-4 and 25253/16). We confirmed the social form of each colony using the (Krieger & Ross 2002) assay on a pool of DNA from 10 randomly chosen workers. Colonies were kept for 6 weeks under the same conditions (natural light, room temperature, cricket, mealworm and honeywater diet) before sampling. From each colony, we snap-froze (between 12:00 and 15:00 local time) one worker and one unmated queen for gene expression analysis. On dry ice, we separated each queen into three body parts using clean (ethanol + bleach) tweezers: head, thorax and abdomen to partially resolve the allometry issue. In total, we had 24 samples for RNA extraction: 6 replicates of 3 body parts from queens and 6 whole-body workers.

We extracted RNA and DNA from each sample using a dual DNA/RNA Tri Reagent based protocol (Annex I, Text AI.1). We applied the (Krieger & Ross 2002) assay on the extracted DNA to identify only individuals with the SB/Sb genotype. Once RNA was extracted, cDNA libraries were prepared from total RNA using half volumes of the NEBNext Ultra II RNA Library Prep Kit for Illumina. Raw RNA and the libraries were quality checked on an Agilent Tapestation 2200; library insert size averaged 350 bp. An equimolar pool of the 24 libraries was sequenced on a single lane of an Illumina HiSeq 4000 sequencer, using 150bp paired

reads. The sequencing produced an average of 14,848,226 pairs of reads per sample, with a maximum of 27,766,980 and a minimum of 6,015,662 (additional information including GPS coordinates in Annex I, Table AI.1b).

The reads generated where then quality checked and aligned to the *S. invicta* genome reference (version gnG; RefSeq GCF_000188075.1; Wurm et al. 2011). More details are available in Annex I, Text AI.2.

Identifying fixed SNP differences between SB and Sb males.

To detect allele-specific differences between SB and Sb we first identified SNPs with fixed differences between the SB and Sb variants. Because the patterns of genetic diversity differ between North American and South American *S. invicta* populations (Ross et al. 2007; Ahrens et al. 2005), we estimated allele-specific expression differences in the social chromosome independently for each population. For this we used haploid male ants because they can provide unambiguous genotypes. For the North American population, we identified fixed allelic differences between a group of 7 SB males and a group of 7 Sb males (NCBI SRP017317, Wang et al. 2013).

For the South American population, we sequenced the genomes of 13 SB males and 13 Sb males sampled from across Argentina. More details in Annex I, Text AI.3 and Table AI.2. For each dataset, we identified fixed allelic differences between the group of SB males and the group of Sb males. We first aligned the reads of each sample to the *S. invicta* reference genome (gnG assembly) using Bowtie2 (v2.3.4; Langmead et al. 2009). We then used Freebayes (v1.1.0; Garrison & Marth 2012) to call variants across all individuals (parameters: ploidy = 1, min-alternate-count = 1, min-alternate-fraction = 0.2). We used BCFtools (v1.9; Li et al. 2009) and VariantAnnotation (v1.30.1; Obenchain et al. 2014) to only retain variant sites with single nucleotide polymorphisms (SNPs), with quality value Q greater or equal than 25, and where all individuals had a minimum coverage of 1. To avoid considering SNPs erroneously called from mis-mapped repetitive reads, we discarded any SNP with mean coverage higher than 16 for the North American samples and 12 for the South American samples (the mean coverage distribution per SNP is shown in Annex I, Fig AI.1) and where one or more individuals had less than 60% reads supporting the called allele. This last filtering step also acts to remove SNPs called from reads with sequencing errors. We then extracted only the SNPs located within the supergene (based on the genomic locations from Pracana, Priyam, et al. 2017) and with a fixed difference between SB and Sb in each population.

We filtered the common SNPs between South and North American populations using BCFtools isec (v1.9; Li et al. 2009). We then used SNPeff (v4.2; Cingolani et al. 2012) to characterise the effect on genes of individual SNPs based on their genomic sequence.

Estimation of read counts from alternate social chromosome variants in heterozygous individuals

Because the reference genome for *S. invicta* is based on an SB individual, read mapping could be biased towards the SB variant in heterozygous individuals, resulting in apparent overall expression of the reference variant (Castel et al. 2015). To overcome this potential artifact, we performed two alignments using STAR (with the same parameters as described in the 'Generation of RNAseq data' section): one with the regular SB reference genome and another with a modified reference genome in which we had replaced the fixed positions between variants in the supergene by those found in Sb. The modified Sb reference was generated using BCFtools consensus (v1.9; Li et al. 2009); this was done once for the North American and once for the South American population. The bam files from alignment to regular and to the modified reference were then merged together using SAMtools (v1.9; Li et al. 2009). Most reads will have been mapped to both references, and thus appear twice in the merged aligned file. These duplicates were removed randomly using rmdup from WASP (van de Geijn et al. 2015) to generate reference bias free alignment files.

We obtained allele-specific counts using GATK's 'ASEReadCounter' (v 3.6-0-g89b7209; McKenna et al. (2010), more information in Annex I, Text AI.4), using the fixed SNPs differences between Sb and SB. Both the Wurm et al. (2011) and Argentinian datasets were run independently in GATK. We imported the resulting variant-specific SNP read counts per sample into R (v3.4.4; R Core Team 2017). We then intersected the counts per SNP with the positions of genes using the R packages 'GenomicRanges' (v1.26.4; Lawrence et al. 2013) and 'GenomicFeatures' (v1.26.3; Lawrence et al. 2013) along with the NCBI protein-coding gene annotation for *S. invicta* (gnG assembly, release 100). The expression of long genes with several fixed SNPs between variants could be overestimated if the reads per SNP are counted individually. To avoid this, we estimated the total expression level for a particular allele (i.e. the SB or Sb variant for any given gene) as the median of all SNP-specific read counts per gene and per variant. For instance, take a given gene with 3 fixed SNPs between SB and Sb. The SB variants for these SNPs would have 12, 15 and 18 reads mapped, and the Sb variants, 5, 8 and 6. In this particular case, we would report that the SB variant for this particular gene has an expression level of 15 reads and the Sb variant, 6 reads.

Expression differences between the SB and Sb variants

We imported the estimated read counts generated by Kallisto into R using Tximport (v1.2.0; Soneson et al. 2015) and DESeq2 (v1.14.1; Love et al. 2014). For subsequent allele-specific expression analyses we filtered out any gene that had less than 3 read counts mapping to either allele. Normalisation methods such as RPKMs could not be used in this case because our methods measured read counts aligning to either variant of each SNP. Additionally, due to the methodology used to obtain allele-specific expression, duplicate reads were removed from the RNAseq dataset, which results in an artificially low read count overall. As a result, normalisation methods based on the total size of the library vastly over-estimated allele-specific expression levels.

Because allele-specific counts come from the same libraries, the standard normalisation methods based on the assumption that most genes are not differentially expressed (Dillies et al. 2013) would artificially reduce real differences in allele expression. For this reason, as recommended by the developers of DESeq2 (Love, 2017), we deactivated normalisation by setting SizeFactors=1. For the Wurm et al. (2011) dataset, we only considered genes expressed in all samples for downstream analyses, whereas for the Argentinian populations RNA dataset, we only analysed genes expressed in all replicates in at least one of the body parts.

To increase the robustness of the expression analyses between the SB and Sb variants of the supergene region, the RNAseq datasets from South American populations (extracted from the Argentinian dataset) and North American populations (from Wurm et al. (2011) we first analysed them together. The South American dataset includes body part information, which is absent in the North American dataset. If both datasets were analysed together with any of the standard tools for gene expression analysis, the effect in expression levels arising from different body parts would be confounded with that arising from differences between the two datasets. To overcome this issue, we performed a linear mixed effects model on the logarithm 2 of the fold change differences between SB and Sb across populations and body parts, using the R packages lme4 (v1.1-18.1; Bates et al. 2014) and lmerTest (v3.1-0; Kuznetsova et al. 2017). The aim of this model was to identify the expression differences between SB and Sb across the different datasets, accounting for the proportion of variance explained by body part and population. We fitted the logarithm 2 fold differences using a 0 intercept with gene, population and their interaction as fixed effects, and the interaction between gene and body part as random effects. Additionally, the logarithm 2 fold differences were weighted by a function of the total read counts per gene. Here we only report the

results of the fixed effects per gene after adjustment of the p values using the Benjamini and Hochberg method (Benjamini & Hochberg 1995). For this joint analysis we only used the genes that had fixed differences between SB and Sb in both South and North American populations.

We also analysed the differential expression of alleles at the supergene region in each population independently. This simpler design allowed us to use more standard tools for the allele-specific expression such as DESeq2 as suggested by Castel et al. (2015). More information in Annex I, Text AI.5.

Expression differences between social forms

We determined the expression levels for all samples from the Morandin et al. (2016) data by using the count mode in Kallisto (v0.44.0; Bray et al. 2016) using *S. invicta* coding sequences (gnG assembly, release 100) as a reference. We imported the estimated counts into DESeq2 (v1.14.1; Love et al. 2014) using Tximport (v1.2.0; Soneson et al. 2015). We compared the DESeq2 normalised expression levels between social forms, determining significance of differential expression using the default Wald test for pairwise comparisons between genes. For each caste we estimated the proportion of significantly differentially to non-differentially expressed genes within and outside the supergene region. The positions of all genes within the supergene region was obtained from (Pracana, Priyam, et al. 2017) and implemented using the R packages GenomicRanges and GenomicFeatures (Lawrence et al. 2013). Out of the 15,058 genes present in the annotation file of *S. invicta*, 10,182 were reliably placed within or outside the supergene region. Only this subset of genes was analysed for testing the enrichment of socially biased loci in the supergene.

Expression differences between variants and social forms

To test whether loci significantly differentially expressed between the SB and Sb variants were also differentially expressed between social forms, we compared the differences identified between social forms (logarithmic fold differences using the Morandin et al. (2016) dataset) with the differences identified between social chromosome variants in the Wurm et al. (2011) dataset. We used a Kolmogorov-Smirnov test to compare the distributions of logarithm fold differences between the SB and Sb variants for the loci more highly expressed in queens from either single or multiple queen colonies. To ensure a balanced comparison, highly expressed genes in single queen colonies were defined as those in the top 50% of logarithmic fold differences for the comparison between social forms (and vice-versa for

genes highly expressed in queens from multiple queen colonies), where positive logarithmic fold differences values indicate high expression in individuals from single queen colonies. To test for a potential enrichment of SB bias genes when analysing genes with differences between variants but not between social forms, we performed a binomial test comparing the number of genes with higher expression in SB. Additionally, we tested whether the overall expression of these genes was biased towards SB by measuring the median log2 fold differences between variants. Genes with higher expression towards SB have positive log2 fold differences. If there is an overall trend towards SB, the median of all genes with differences between variants only should be significantly different from 0. We tested the significant deviation from 0 through a Wilcoxon signed rank test.

Finally, we also explored whether genes with low expression of their Sb allele showed increased expression of their SB allele, resulting in similar expression between multiple-queen (SBSb genotype) and single-queen (SBSB genotype) individuals. Such a pattern would be consistent with an ongoing process of dosage compensation. To do so, we excluded genes with significant biases towards Sb and high SB/SBSb ratios (*i.e.* SB variant more highly expressed in SBSb than SBSB individuals). The rest of all analysed genes were then grouped by relative SB/Sb expression. We then compared the overall expression levels between these groups in multiple-queen and single-queen individuals. In case of mechanisms such as dosage compensation we would expect that the overall expression levels between social forms in each group should remain similar despite SB-Sb expression differences.

Because the differences between social forms calculated from the Morandin et al. (2016) dataset were based only on North American populations, we did not repeat this comparison with the Argentinian dataset. As discussed above, both populations have different levels of genetic diversity, and the genes involved in social form differences are not necessarily the same. Comparing the expression differences between variants and social forms without accounting for the differences between the North and South American populations would result in confounding factors underlying the observed results.

Results

Fixed differences between supergene variants

In order to perform an analysis of the expression differences between supergene variants, we first identified SNPs which are different between variants and fixed in either SB or Sb across South and North American populations of *S. invicta*. We found that the two populations had 2,877 SNPs between variants in common.

These variant-specific SNPs in common across populations affected 352 genes within the supergene region. The vast majority of SNPs fell in intronic regions (57.96%), only 3.43% of the variants affected exonic regions and 1.99% and 0.54% affected 3’ and 5’ UTR regions respectively. Of the variants that had an impact in coding sequences (222 in total), we found that slightly more SNPs Sb (52.25%) resulted in silent mutations, than in missense mutations (47.3%) and only one SNP resulted in a nonsense mutation (0.5%).

In all, the vast majority of SNPs in Sb were predicted to have potential regulatory effects (96.56%), a few variants could disrupt protein effectiveness (1.51%) or instead, have limited or no impact on protein structure (1.91%). Only two Sb variants (0.45%) were predicted to have a high impact on protein function compared to SB.

Within each population, we identified 3,129 of such differences in the South American populations (92% of which are in common with North American populations), and 25,899 in North American (11% of which are in common with South American populations). The greater number of fixed differences in North America is likely caused by the bottleneck during the introduction of this species in the 20th century, that led to a strong drop in genetic diversity (Ross & Shoemaker 2008).

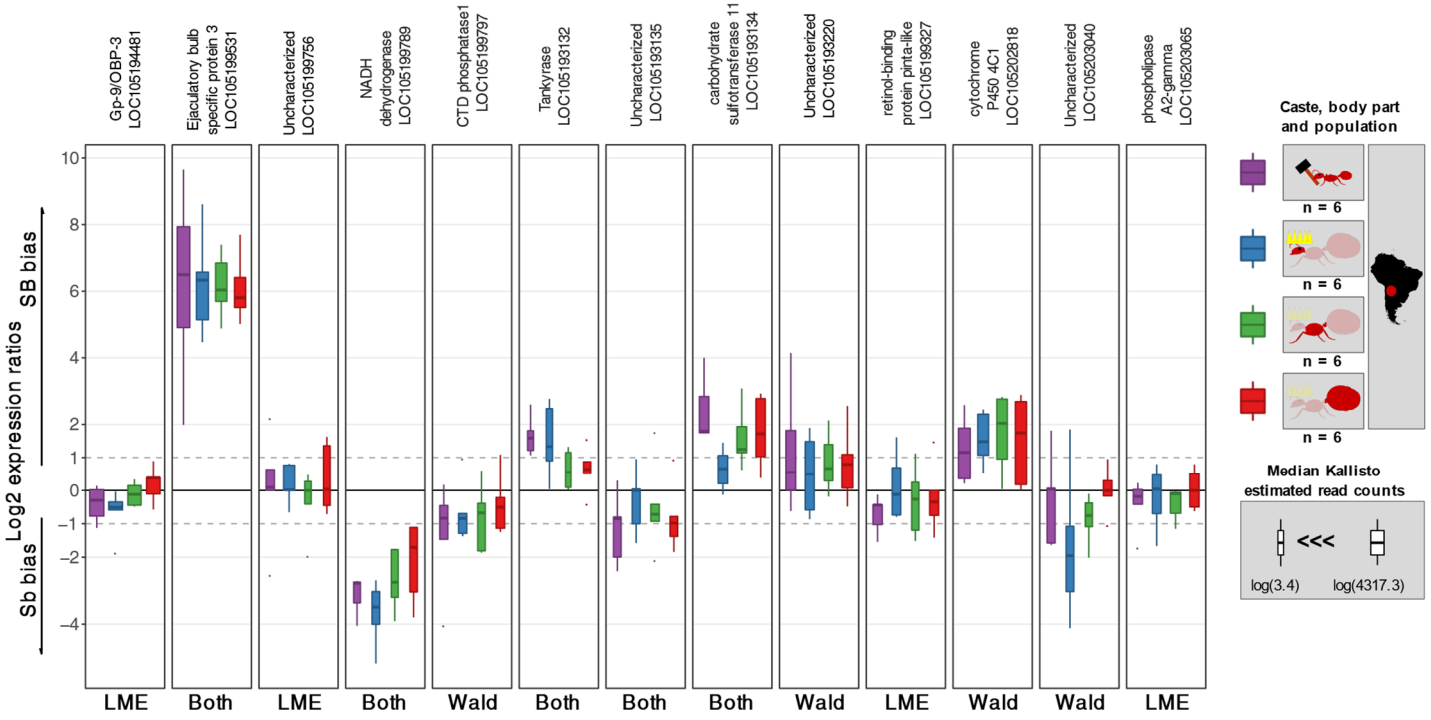


Figure 2.1. Allele-specific expression for genes in the fire ant social supergene. Differences in expression (y axis) between the variants of the social chromosome in South American workers (whole body), queen heads, queen thoraces or queen abdomens. We show relative expression levels for social chromosome genes that had fixed SNP differences between SB and Sb in both populations as well as significantly expression differences. The text under each plot indicates which test showed a significant (Benjamini and Hochberg corrected p value < 0.05) difference between variants for both populations: LME, linear mixed effects model including North American populations; Wald, Wald test from DESeq2 performed on RNAseq from South American populations only or; Both, the two tests showed significant differences in expression between variants. Within each plot, each box shows the distribution of logarithm 2 fold differences between SB and Sb per caste/body part. Genes with values above the solid line (logarithm 2 fold differences = 0) are more highly expressed in SB and vice versa.

Supergene variant-specific expression patterns

Because the vast majority of coding sequences are intact in both variants of the supergene, we asked whether variant-specific expression biases may occur. For this, we obtained RNAseq gene expression data from SB/Sb individuals and identified fixed genetic differences between SB and Sb individuals. Specifically, we generated new individual-specific RNA-seq data from whole bodies of SB/Sb workers and from abdomens, thoraces, and heads of virgin queens collected in South America and additionally, we analysed existing RNA-seq gene expression data from pools of whole bodies of SB/Sb queens collected in the USA (Wurm et al. 2011). We then used the genes with fixed differences between variants identified previously to estimate allele-specific expression differences between SB and Sb.

Among the 352 genes with fixed differences between SB and Sb, 61 had sufficient expression for analysis of differences between alleles. Within populations, we found 69 such genes in the native South American range and 213 such genes in the invasive North American population. Most of the genes with variant-specific expression in South American populations (88.4%) were also found in the populations of the invasive range.

We then tested whether any of the genes with fixed differences between SB and Sb in both populations showed differences in expression between variants. To do so, we ran a linear mixed effects using the RNAseq data from both South and North America. According to the model, 9 of the 61 genes had consistent allele-biased expression (Fig 2.1, Annex I, Table AI.3): the Sb variant was consistently more highly expressed for ‘pheromone-binding protein Gp-9/OBP3’ (LOC105194481), ‘retinol-binding protein pinta-like’ (LOC105199327), ‘NADH dehydrogenase’ (LOC105199789) and ‘uncharacterized gene’ (LOC105193135), while the SB variant was more highly expressed for ‘ejaculatory bulb-specific protein 3’ (LOC105199531), ‘carbohydrate sulfotransferase 11-like’ (LOC105193134), ‘calcium-independent phospholipase A2-gamma’ (LOC105203065), ‘Tankyrase’ (LOC105193132) and ‘uncharacterized gene’ (LOC105199756). Four of these nine genes with variant-specific differences (Annex I, Table AI.3) also show differences between social forms: “pheromone-binding protein Gp-9” (LOC105194481), ’retinol-binding protein pinta-like‘ (LOC105199327), ‘NADH dehydrogenase’ (LOC105199789) and ’ejaculatory bulb-specific protein 3‘ (LOC105199531) were all more highly expressed in queens from multiple-queen colonies than queens from single-queen colonies. These genes had respectively 20, 4, 1 and 3 fixed SNPs between SB and Sb in both South and North American populations. None of the SNPs were located in exonic regions and the vast majority of SNPs were predicted to have non-

coding effects or effects unlikely to change protein behaviour. The only exception were 4 variants in “pheromone-binding protein Gp-9”, which were predicted to have a potential effect in protein effectiveness. More details in Annex I, Fig AI.2.

To potentially increase power by eliminating geographical structure in the data, we additionally analysed each population independently using a DESeq2-based approach (v1.14.1; Love et al. 2014). All of the 9 previously identified genes also showed allele-biased expression using this approach. We additionally identified 9 genes with allele-biased expression exclusive to the South American data (Fig 2.1), and 23 genes with allelle-biased expression differences exclusive to the North American data (Annex I, Fig AI.3). None of the genes with North American variant-specific variation had fixed differences in South American populations, and were therefore excluded from the joint analysis. Because the South American RNAseq data includes body part and caste information (head, thorax and abdomen in queens and whole body in workers), we tested whether differences in expression between SB and Sb varied across body parts and castes. We found no significant interaction effect for any of the genes (DESeq2’s Logarithmic Ratio Test adjusted p value > 0.05 for all genes analysed. Further in line with this, all pairwise comparisons between interaction terms had BH adjusted p values > 0.05 for all genes analysed). More details regarding the independent analyses are available in Annex I (Text AI.6, Table AI.4 and Fig AI.3).

The social supergene is enriched in socially biased loci in queens

We identified which genes are differentially expressed between social forms (socially biased genes) and whether they are more common than expected in the supergene region. To do so, we compared RNAseq data from whole bodies of egg-laying queens from single and multiple-queen colonies (Morandin et al. 2016).

We identified 426 socially biased genes (the complete list of genes is available in Online AnnexI at http://bit.ly/OnlineAI), from which we established chromosomal location for 343. Among these socially biased genes, genes in the supergene region were significantly overrepresented (Fig 2.2a, 33 out of 343 differentially expressed genes, 12 expected by chance, $\chi^2 = 29.7$, p value $< 10^{-7}$). Additionally, we found that the vast majority of socially biased genes (400 out of 426, *i.e*. 94%) were more highly expressed in multiple-queen colonies than in single-queen colonies (binomial test, p value $< 10^{-7}$, Fig 2.2b). We obtained independent confirmation of this result by comparing expression of queens from multiple-queen colonies from another RNAseq dataset (Wurm et al. 2011) to those of queens from single queen colonies from Morandin et al. (2016) (Annex I, Text AI.7).

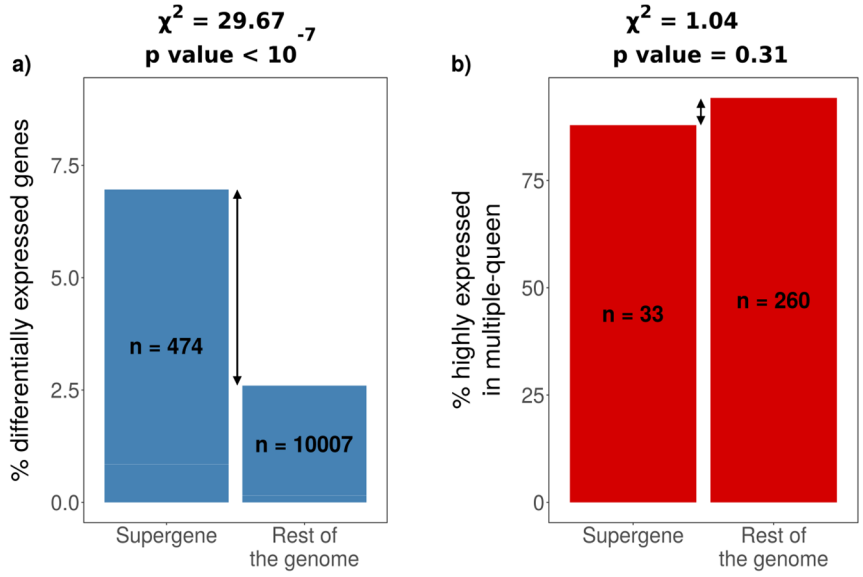


Figure 2.2. Distribution of the proportion of differentially expressed genes within (left bar) and outside (right bar) of the supergene region. Within each bar, 'n' indicates the total number of genes in which the proportions are based. **a)** differential expression between queens from single-queen and multiple-queen colonies. **b)** the proportion of significantly differentially expressed genes that are more highly expressed in multiple-queen colonies. The vast majority of genes with differential expression between social forms are more highly expressed in multiple-queen colonies. The proportion of genes being more highly expressed in multiple-queen colonies is similar within and outside the supergene.

Loci with higher expression in Sb are also more highly expressed in queens from multiple-queen colonies.

To better understand the potential evolutionary forces that may be shaping expression patterns within the social chromosome we performed a joint analysis comparing the results of differences in expression between variants (based on the Morandin et al. (2016) data) and between social forms (Wurm et al. 2011). Using expression as a proxy for benefit, under evolutionary conflict we expect that genes which Sb variant is more highly expressed to be also more highly expressed in multiple-queen individuals. On the other hand, we expect that if Sb is accumulating deleterious mutations, genes affected by deleterious Sb alleles will be downregulated in this variant (and therefore upregulated in SB) and will show no differences in expression between social forms due to dosage compensation.

We found 8 genes with both allele-biased and socially-biased expression. There was a strong directional bias, where 5 of the genes were more highly expressed in Sb and in multiple-queen colonies (Fig 2.3d). This trend, however, was not significant after subsampling our dataset 1000x to equate the number of multiple-queen and single-queen biased genes. We also found that 14 genes were only differentially expressed between variants but not between social form (Fig 2.3c). Of these, 11 were more highly expressed in SB, a marginally non-significant difference (binomial test, p value = 0.057). The median fold difference between SB and Sb for these 14 genes, however, was 1.7, indicating significant bias in general expression towards SB (Wilcoxon signed rank test p value = 0.013). This pattern is consistent with ongoing mechanisms of dosage compensation arising from Sb degeneration, whereby the low expression of the Sb allele would be rescued by higher expression in SB. As a result, the expression level for these genes would be similar in both single-queen (SB/SB) and multiple-queen (SB/Sb) individuals. To further investigate this possibility, we compared the relative expression levels of the SB and Sb alleles in queens from multiple queen-colonies to the total expression of the genes in the supergene in either social form (Fig 2.4). The figure shows that as the differences between SB and Sb increase, the differences in expression between single-queen and multiple-queen individuals remain non-significant (Wilcoxon between the overall expression differences for each group of SB/Sb relative expression p value > 0.05).

Most genes within the supergene showed no differential expression patterns. Out of the 202 genes for which we had both social form and variant-specific gene expression information, 171 (85%) showed no significant differences between social forms nor between variants (Fig 2.3a), while 9 genes (4%) were only differentially expressed between social forms. The latter category of genes also showed a strong bias towards multiple-queen higher expression (Fig 2.3b).

Discussion

Sb expression patterns are consistent with adaptation to multiple-queen form

Our results based on RNAseq data from queens of *S. invicta* from populations in the North American invasive range show patterns which are consistent with adaptive processes influencing the evolution of the social chromosome. We found that the supergene region is enriched in genes which are differentially expressed between social forms. Additionally, we show that genes that were more highly expressed in the Sb variant, tend to also be more highly expressed in queens from multiple-queen colonies, the phenotype with which this variant is fully linked. This is consistent with the idea that at least some alleles in the Sb variant are involved in defining the multiple-queen phenotype. If we take expression levels as a proxy for benefit, we should expect that the alleles which show differential expression between phenotypes within the supergene variant that spends more time in a given phenotype should be more highly expressed in that phenotype. Our results show these same expression patterns in the *S. invicta* supergene, expected to emerge from evolutionary conflict. These patterns are analogous to those seen in other supergene systems (e.g. Sun et al. 2018) in general, and sex chromosomes in particular (Wright et al. 2017; Ellegren & Parsch 2007; Lipinska et al. 2015).

We do not find, however, a similar enrichment for SB expression in single-queen colonies. A potential explanation for this could be linked to the complex selective pressures at play in SB. Because Sb is only found in multiple-queen colonies, selection would always push towards the multiple-queen optimum in Sb. The selective forces acting on SB, however, are more complex to predict because this variant spends a variable amount of time in both phenotypes. To make matters more complex, the selective pressures will also act during the long haploid phase of the reproductive cycle of ants (Hall & Goodisman 2012). Similarly to what happens in some X chromosomes in XY systems (Patten 2019), we do not necessarily expect SB to be enriched in single-queen colony alleles.
It is also important to note that there is a strong expression bias towards queens from multiple-queen colonies, in general, we find that genes which are differentially expressed between social forms are more highly expressed in queens from multiple-queen colonies. This bias in expression could be making it difficult to detect a pattern in SB similar to that found in Sb, simply because there are not enough single-queen biased genes to make the pattern detectable. Indeed, after subsampling 1000x we did not find any enrichment in Sb compared to SB, which is consistent with a lack of power to detect SB enrichment of single-queen colony biased genes.

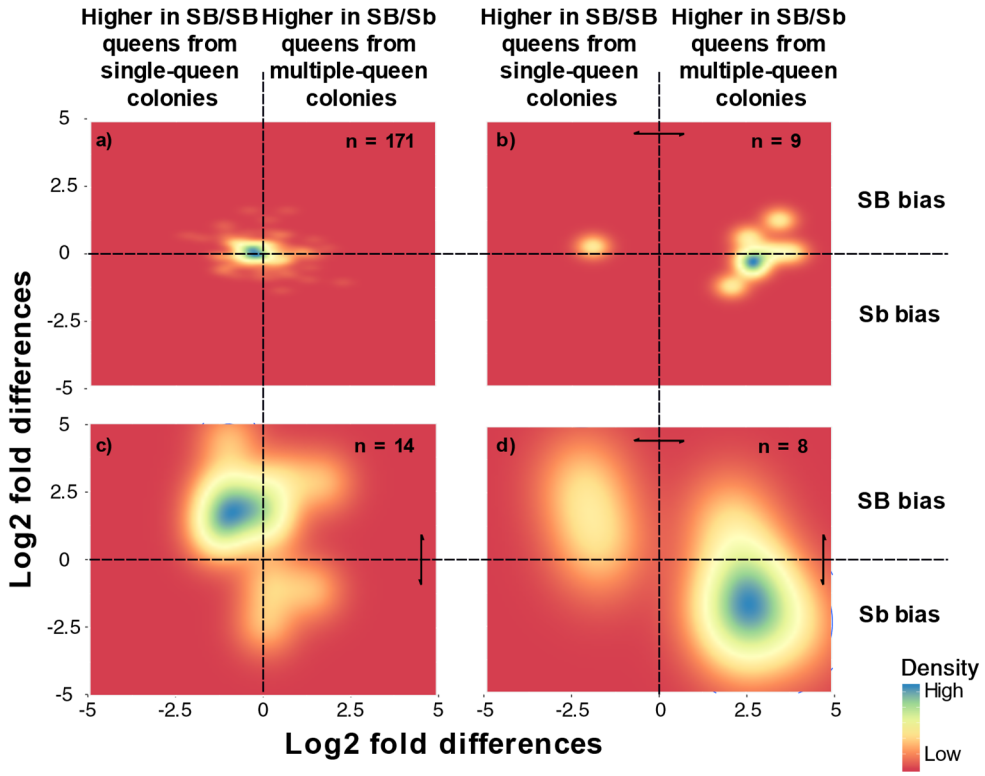


Figure 2.3. Distribution of loci showing different patterns of expression. The density plots represent the number of loci showing each pattern of expression. The X-axis of each plot gives the ratio of a locus's expression in SBSb queens relative to SBSB queens (log scale). The Y-axis gives the ratio of that locus's expression of Sb alleles relative to SB (in SBSb queens). Panel a) summarizes the results from loci showing no significant difference in either ratio, which therefore cluster around the (0,0) origin. The remaining three panels summarise the patterns for other loci: b) those loci only showing significant differences between queens (i.e. SBSb *vs* SBSB); c) those loci only showing differences between alleles (i.e. SB vs Sb); d) loci showing both types of bias. Notice that in each of panels b, c and d there is an unequal distribution of loci on either side of the origin. The higher density in the lower-right quadrant of panel d illustrates a pattern in which loci that show high relative expression of the Sb allele in heterozygotes ($x > 0$) tend to be more highly expressed in queens from multiple-queen colonies ($y < 0$). This pattern, however, could not be replicated after subsampling the number of multiple-queen biased genes 1000x, which suggest that this study is under-powered to detect a similar enrichment of multiple-queen biased genes in SB Comparisons between expression levels in queens were made using data from (Wurm et al. 2011) and (Morandin et al. 2016); the allelic expression are reported for the first time in this paper.

Expression patterns are consistent with markers of non-adaptive processes

We generated a new RNAseq dataset with different body parts of queens and whole-body workers from the native South American range of *S. invicta* to compare variant-specific expression in the supergene across different populations, body parts and castes. The patterns we found here suggest that most expression patterns within the supergene are driven by processes arising from the lack of recombination. If the Sb alleles were performing overall functions fundamentally different than its SB counterpart, we would expect high levels of differential expression between variants. In addition, those expression differences between variants should be different across body parts and castes.

Within the South American population, we find that the vast majority of genes do not show differential expression between supergene variants. Additionally, we find that the few genes being differentially expressed between variants, show the same patterns of expression across body parts and castes. Similar results have been found in other supergene systems (e.g. Sun et al. 2018). Allele-specific expression is expected to vary across body parts (The GTEx Consortium 2015). This pattern could indicate that most of the expression differences observed between SB and Sb are due to the low diversity of the genetic environment in the supergene region. Consequently, most expression differences observed between supergene variants would not necessarily play an active role in defining the phenotype with which the supergene is linked, in this case, social form.

To further this argument, we also find many more differences between variants in the North American populations of *S. invicta* than in the Southern ones. This is because for our analyses we only took into account genes with fixed differences between SB and Sb, and since genetic diversity is lower in North American populations (Ahrens et al. 2005), we found more genes with fixed differences between variants in those populations. Most of the significant differences between variants in North American populations affected genes that do not have fixed differences in South American populations. This indicates that genes with expression differences between variants only in North America are not necessary to explain differences in social forms.

From these results, we conclude that many of the expression differences that we observed between variants may have been caused by factors arising from the lack of recombination in Sb, that are not necessarily involved in defining social phenotype. For instance, in addition to the expression data, we also find that of all the fixed mutations between SB and Sb, roughly

half are silent (synonymous) and half are missense/nonsense (non-synonymous). This finding is in line with the increased ratio of non-synonymous to synonymous mutations in the supergene region relative to the rest of the genome reported by (Pracana, Priyam, et al. 2017). Additionally, Sb is enriched in transposable elements and other repetitive elements due to inefficient purifying selection (Stolle et al. 2019). New insertions in the promoter regions of the Sb variant could cause changes in expression levels of any Sb allele (Cowley & Oakey 2013). The change in expression could then be maintained even if it had mildly deleterious fitness effects in Muller's ratchet-type of process (Bachtrog 2013).

Expression patterns are consistent with dosage compensation in degenerating Sb alleles

Variant-specific differences in expression between loci within the supergene could also be explained by some form of dosage compensation. Inefficient purifying selection can result in an accumulation of deleterious mutations in the alleles trapped in the Sb variant. Indeed, higher levels of non-synonymous mutations have been reported in Sb (Pracana, Priyam, et al. 2017) and would be, consequently, downregulated. If those particular genes need to be expressed to a specific level (e.g. housekeeping genes), the 'healthy' alleles in SB for those genes would compensate for the lack of Sb expression. We therefore explored the possibility of dosage compensation in the Sb variant at a gene by gene level, as seen in other young supergene systems (Nozawa et al. 2014; Alekseyenko et al. 2013; Sun et al. 2018). Many of the genes in Sb seem to be accumulating deleterious mutations (in coding and/or regulatory regions) (Pracana, Priyam, et al. 2017). We found that most genes with differences between variants that do not show expression differences between social forms tend to be more highly expressed in SB. That is, whenever we find differences between variants in heterozygote individuals, the total dosage level for those particular genes is similar to homozygote individuals. In most cases, those genes have an SB-biased expression. This does not necessarily mean that all the genes with higher expression in SB are affected by dosage compensation, but it indicates that most of them are affected, as to result in a bias for the overall expression patterns in the social chromosome. Additionally, we found that, overall, changes in the ratio of expression between the SB and Sb variants does not result in differences in expression between social form, further supporting the notion of dosage compensation acting on degrading Sb alleles. This finding is similar to that found in other supergene systems in very different organisms such as the dioecious plant *Silene latifolia* (Muyle et al., 2012).

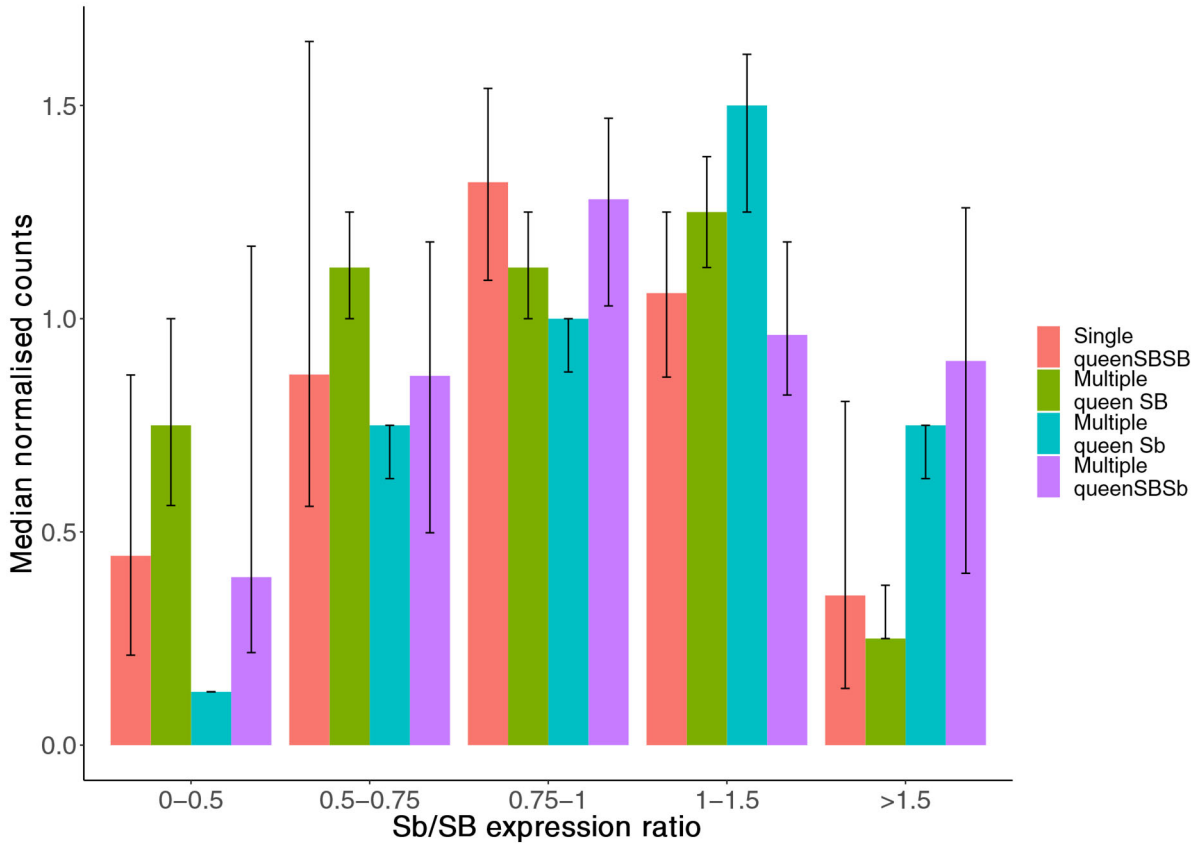


Figure 2.4. Expression differences between single-queen (red bars) and multiple-queen (purple bars) individuals for different levels of SB (green bars) to Sb (blue bars) expression in multiple-queens. The plot represents the overall expression levels for all genes analysed in the supergene for which there was both allele-specific and social form expression data. Genes with significant expression biases towards Sb or with a high SB/SBSb ratio were treated as outliers and removed. As a result, 193 genes in total were included in this analysis. Each bar represents the median normalised expression within group, the error bars are the 95% CI interval around the median (estimated from a 5000x median bootstrapping). The expression levels within each expression group (Single-queens, multiple-queens, SB and Sb expression) is normalised by the total number of reads in that group. This normalisation allows the comparison across different datasets with different levels. The differences in expression between single-queen and multiple-queen individuals remains non-significant (Wilcox test p value > 0.05) across varying levels of SB-Sb expression differences. Only when Sb expression is much larger than that of SB (Sb/SB expression ratio > 1.5) does it seem to be an increase in gene expression for multiple-queen individuals, but the difference between social forms remained non-significant (Wilcox test p value = 0.08). These results are consistent with SB expression compensating for low Sb expression.

Candidate genes for differences between social forms

We have hypothesised that most expression differences observed between the SB and Sb variants of the *S. invicta* supergene are not necessarily linked to differences in social form, while acknowledging at the same time that some of the expression differences are likely to be playing a role in social polymorphism. Hypothesising which genes and exactly which role they are playing in the definition of social form is beyond the scope of this study, but based on our results we can single out candidate genes that may be involved in the social polymorphism of *S. invicta*. We found seven genes with variant-specific gene expression differences in both South and North American populations. As we have seen above, these differences in expression are not sufficient to link those genes to differences in social form. We find that three such genes are also differentially expressed between social forms in North American populations, which adds another layer of evidence pointing at a potential role in social form differences for these genes. Namely, “pheromone-binding protein Gp-9”, “ejaculatory bulb-specific protein 3” and “retinol-binding protein pinta-like”. “Pheromone-binding protein Gp-9” was one of the first genetic markers used to identify multiple- or single-queen colonies in *S. invicta* (Ross 1997). It has been identified as an odorant binding protein, more specifically, OBP-3 (Pracana, Levantis, et al. 2017). This makes it an interesting candidate, since its link with social form as a green-beard gene has been hypothesised for decades (Keller & Ross 1998), particularly due to its function as odorant protein, which could play a role in ant behaviour at the colony level (Trible et al. 2017). We found 20 fixed differences between SB and Sb for this gene, 4 of which could have an impact on protein efficiency, and 15 with a potential effect on its regulation. “Ejaculatory bulb-specific protein 3” is more highly expressed in SB, it is also identified as an insect odorant binding protein (InterPro id IPR005055). It has been associated with several functions in *Drosophila melanogaster* including mating (Laturney & Billeter 2014) or viral response (Sabatier et al. 2003). It has also been linked to sexual behaviour in the moth *Mamestra brassica* (Bohbot et al. 1998), subcaste differences in the bumblebee *Bombus impatiens* (Wolschin et al. 2012), venom production in social hornets (Yoon et al. 2015) and caste differences in the termite *Reticulitermes flavipes* (Steller et al. 2010). We found 3 fixed differences between SB and Sb for this gene, one of which is in the 3’ end UTR, with a potential effect on gene regulation. NADH-dehydrogenase plays a role in electron transport, and, consequently, in regulating metabolic rates. We found 1 fixed difference between variants for this gene. Finally, “ retinol-binding protein pinta-like” is similar to the PINTA retinol-binding protein, which is linked to pigment transport and vision in *D.melanogaster* and the butterfly *Papilio xuthus* (Pelosi et al. 2018). Interestingly, nearly all genes that we

identified as potential candidates are involved in environmental perception, and at least one of them has been linked to caste differences in other social insects. We found 4 fixed differences between variants for this gene, almost all of which are predicted to impact gene regulation.

Conclusions

We analysed RNAseq data from different populations, castes and body parts of the fire ant *Solenopsis invicta*. We found that evolutionary conflict accounts for part of the expression patterns observed within the supergene region. Specifically, 1) an enrichment on genes differentially expressed between social form within the supergene region and 2) an enrichment in the Sb variant of genes more highly expressed in multiple-queen colony queens. We also find, however, that most expression patterns within the social chromosome are most likely due to non-adaptive processes due to suppressed recombination. We reach this conclusion because 1) most genes are not differentially expressed between variants, 2) the few genes that show variant-specific differential expression tend to not be involved in social form differences, and such genes are 3) more frequently highly expressed in SB, suggesting gene-specific dosage compensation is at play, additionally, 4) many genes displaying variant-specific expression in North America do not have fixed differences between SB and Sb in South American populations. Overall, these results show that evolutionary conflict is likely to have driven the suppression of recombination initially, but that non-adaptive processes are responsible for most of the expression patterns observed.

References

Ahrens, M.E., Ross, K.G. & Shoemaker, D.D., 2005. Phylogeographic structure of the fire ant *Solenopsis invicta* in its native South American range: roles of natural barriers and habitat connectivity. *Evolution; international journal of organic evolution*, 59(8), p.1733–1743.

Alekseyenko, A.A. et al., 2013. Conservation and de novo acquisition of dosage compensation on newly evolved sex chromosomes in *Drosophila*. *Genes & development*, 27(8), p.853–858.

Bachtrog, D., 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature reviews. Genetics*, 14(2), p.113–124.

Bates, D. et al., 2014. Fitting Linear Mixed-Effects Models using lme4. arXiv arXiv1406.5823

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 57(1), p.289–300.

Bergero, R. & Charlesworth, D., 2009. The evolution of restricted recombination in sex chromosomes. *Trends in ecology & evolution*, 24(2), p.94–102.

Bohbot, J. et al., 1998. Functional characterization of a new class of odorant-binding proteins in the moth *Mamestra brassicae*. *Biochemical and biophysical research communications*, 253(2), p.489–494.

Branco, S. et al., 2017. Evolutionary strata on young mating-type chromosomes despite the lack of sexual antagonism. *Proceedings of the National Academy of Sciences of the United States of America*, 114(27), p.7067–7072.

Branco, S. et al., 2018. Multiple convergent supergene evolution events in mating-type chromosomes. *Nature communications*, 9(1), p.2000.

Bray, N.L. et al., 2016. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5), p.525–527.

Castel, S.E. et al., 2015. Tools and best practices for data processing in allelic expression analysis. *Genome biology*, 16, p.195.

Cavoto, E. et al., 2018. Sex-antagonistic genes, XY recombination and feminized Y

chromosomes. *Journal of evolutionary biology*, 31(3), p.416–427.

Charlesworth, D., 2016. The status of supergenes in the 21st century: recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. *Evolutionary applications*, 9(1), p.74–90.

Cingolani, P. et al., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(2), p.80–92.

Cowley, M. & Oakey, R.J., 2013. Transposable elements re-wire and fine-tune the transcriptome. *PLoS genetics*, 9(1), p.1003234.

Darlington, C.D. & Mather, K., 1949. *The elements of genetics*, George Allen & Unwin Ltd: London.

Dillies, M.-A. et al., 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*, 14(6), p.671–683.

Dufresnes, C. et al., 2015. Sex-chromosome homomorphy in palearctic tree frogs results from both turnovers and X–Y recombination. *Molecular biology and evolution*, 32(9), p.2328–2337.

Ellegren, H. & Parsch, J., 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nature reviews. Genetics*, 8(9), p.689–698.

Ellison, C. & Bachtrog, D., 2018. Recurrent gene amplification on *Drosophila* Y chromosomes suggests cryptic sex chromosome drive is common on young sex chromosomes. *bioRxiv*

Fritz, G.N., Vander Meer, R.K. & Preston, C.A., 2006. Selective male mortality in the red imported fire ant, *Solenopsis invicta*. *Genetics*, 173(1), p.207–213.

Garrison, E. & Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. arXiv arXiv-1207.3907.

van de Geijn, B. et al., 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature methods*, 12(11), p.1061–1063.

Gotzek, D. & Ross, K.G., 2007. Genetic regulation of colony social organization in fire ants: an integrative overview. *The Quarterly review of biology*, 82(3), p.201–226.

Gu, L. & Walters, J.R., 2017. Evolution of sex chromosome dosage compensation in animals: a beautiful theory, undermined by facts and bedeviled by details. *Genome Biology and Evolution*, 9(9), p.2461–2476.

Hall, D.W. & Goodisman, M.A.D., 2012. The effects of kin selection on rates of molecular evolution in social insects. *Evolution; international journal of organic evolution*, 66(7), p.2080–2093.

Johnson, B.R., Atallah, J. & Plachetzki, D.C., 2013. The importance of tissue specificity for RNA-seq: highlighting the errors of composite structure extractions. *BMC genomics*, 14, p.586.

Keller, L. & Ross, K.G., 1998. Selfish genes: a green beard in the red fire ant. *Nature*, 394, p.573.

Khil, P.P. et al., 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nature genetics*, 36(6), p.642–646.

Krieger, M.J.B. & Ross, K.G., 2002. Identification of a major gene regulating complex social behavior. *Science*, 295(5553), p.328–332.

Küpper, C. et al., 2016. A supergene determines highly divergent male reproductive morphs in the ruff. *Nature genetics*, 48(1), p.79–83.

Kuznetsova, A., Brockhoff, P.B. & Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13).

Langmead, B. et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), p.R25.

Laturney, M. & Billeter, J.-C., 2014. Neurogenetics of female reproductive behaviors in *Drosophila melanogaster*. *Advances in genetics*, 85, p.1–108.

Lawrence, M. et al., 2013. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8), p.1003118.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), p.2078–2079.

Li, J. et al., 2016. Genetic architecture and evolution of the S locus supergene in *Primula vulgaris*. *Nature Plants*, 2(12), p.16188.

Lipinska, A. et al., 2015. Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga *ectocarpus*. *Molecular biology and evolution*, 32(6), p.1581–1597.

Love, M.I., 2017. Using RNA-seq DE methods to detect allele-specific expression. URL: http://rpubs.com/mikelove/ase. Accessed on 22 May 2018

Love, M.I., Wolfgang, H. & Simon, A., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12).

Mank, J.E., 2017. The transcriptional architecture of phenotypic dimorphism. *Nature ecology & evolution*, 1(1), p.6.

McKenna, A. et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9), p.1297–1303.

Morandin, C. et al., 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome biology*, 17, p.43.

Muyle, A. et al., 2012. Rapid de novo evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes. *PLoS biology*, 10(4), p.1001308.

Nozawa, M. et al., 2014. Tissue- and stage-dependent dosage compensation on the neo-X chromosome in *Drosophila pseudoobscura*. *Molecular biology and evolution*, 31(3), p.614–624.

Obenchain, V. et al., 2014. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14), p.2076–2078.

Parsch, J. & Ellegren, H., 2013. The evolutionary causes and consequences of sex-biased gene expression. *Nature reviews. Genetics*, 14(2), p.83–87.

Patten, M.M., 2019. The X chromosome favors males under sexually antagonistic selection. *Evolution; international journal of organic evolution*, 73(1), p.84–91.

Pelosi, P. et al., 2018. Beyond chemoreception: diverse tasks of soluble olfactory proteins in insects. *Biological reviews of the Cambridge Philosophical Society*, 93(1), p.184–200.

Pracana, R., Levantis, I., et al., 2017. Fire ant social chromosomes: Differences in number, sequence and expression of odorant binding proteins. *Evolution letters*, 1(4), p.199–

210.

Pracana, R., Priyam, A., et al., 2017. The fire ant social chromosome supergene variant Sb shows low diversity but high divergence from SB. *Molecular ecology*, 26(11), p.2864–2879.

R Core Team, 2017. R: A language and environment for statistical computing.

Ross, K.G. et al., 2007. Genetic variation and structure in native populations of the fire ant Solenopsis invicta: evolutionary and demographic implications. *Biological journal of the Linnean Society. Linnean Society of London*, 92(3), p.541–560.

Ross, K.G., 1997. Multilocus evolution in fire ants: effects of selection, gene flow and recombination. *Genetics*, 145(4), p.961–974.

Ross, K.G. & Shoemaker, D.D., 2008. Estimation of the number of founders of an invasive pest insect population: the fire ant *Solenopsis invicta* in the USA. *Proceedings of the Royal Society B: Biological Sciences*, 275(1648), p.2231–2240.

Sabatier, L. et al., 2003. Pherokine-2 and-3: Two *Drosophila* molecules related to pheromone/odor-binding proteins induced by viral and bacterial infections. *European journal of biochemistry / FEBS*, 270(16), p.3398–3407.

Soneson, C., Love, M.I. & Robinson, M.D., 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, p.1521.

Steller, M.M., Kambhampati, S. & Caragea, D., 2010. Comparative analysis of expressed sequence tags from three castes and two life stages of the termite *Reticulitermes flavipes*. *BMC genomics*, 11, p.463.

Stolle, E. et al., 2019. Degenerative expansion of a young supergene. *Molecular biology and evolution*, 36(3), p.553–561.

Sun, D. et al., 2018. Rapid regulatory evolution of a nonrecombining autosome linked to divergent behavioral phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), p.2794–2799.

The GTEx Consortium, 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235), p.648–660.

Trible, W. et al., 2017. orco mutagenesis causes loss of antennal lobe glomeruli and impaired social behavior in ants. *Cell*, 170(4), p.727–735.e10.

Vicoso, B., Kaiser, V.B. & Bachtrog, D., 2013. Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16), p.6453–6458.

Wang, J. et al., 2013. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature*, 493(7434), p.664–668.

Wolschin, F. et al., 2012. Size-related variation in protein abundance in the brain and abdominal tissue of bumble bee workers. *Insect molecular biology*, 21(3), p.319–325.

Wright, A.E. et al., 2017. Convergent recombination suppression suggests role of sexual selection in guppy sex chromosome formation. *Nature communications*, 8, p.14251.

Wurm, Y. et al., 2011. The genome of the fire ant *Solenopsis invicta*. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), p.5679–5684.

Yoon, K.A. et al., 2015. Comparative functional venomics of social hornets *Vespa crabro* and *Vespa analis*. *Journal of Asia-Pacific entomology*, 18(4), p.815–823.

Zemp, N. et al., 2016. Evolution of sex-biased gene expression in a dioecious plant. *Nature plants*, 2(11), p.16168.

Chapter 3: A fluctuating gene flow between discrete phenotypes is likely to have maintained the emergence of a supergene

Abstract

The occurrence of several discrete multi-trait phenotypes in a population (complex stable polymorphisms) is a widespread phenomenon in several organisms. On several occasions, the differences between such complex phenotypes are maintained by supergenes. Supergenes are low recombination regions of the genome that maintain tight linkage between two or more loci underlying a complex phenotype. These regions are able to prevent the independent spread of alleles co-adapted to a specific genetic and ecological environment. There are two main evolutionary processes that could result in selection favouring low recombination between loci, leading ultimately to the emergence of supergenes. Namely, antagonistic selection and gene flow between locally adapted populations. It is not well understood the relative role each may be playing in the emergence of supergenes. Here I model the spread of alleles with different fitness effects in different phenotypes. The model simulates the spread of such alleles under different scenarios of antagonistic selection and gene flow between phenotypes. I use the life history traits of the fire ant *Solenopsis invicta* to parametrise the model. This ant species has two types of social organisation: colonies can either have one or multiple queens. This difference in social organisation affects the behavioural and physiological features of the whole colony at is controlled by a supergene. The supergene of *S. invicta* is likely to be subjected to both antagonistic selection and the effects of gene flow between phenotypes, which makes it a relevant model to explore the relative effects of these processes in supergene evolution. The results of the model show that gene flow is very likely the key factor explaining the emergence of the supergene in *S. invicta*. These results shed light on the dynamics of supergene evolution in particular and local adaptation with gene flow in general.

Introduction

The presence of multiple discrete phenotypes in a population is relatively common in the tree of life. The most widespread example is perhaps sexual dimorphism, which can be found in one way or another in groups as diverse as animals (e.g. Shine 1989; Dunn et al. 2001; Desjardins & Fernald 2009), plants (e.g. Sakai & Weller 1999; Barrett & Hough 2013) or algae (e.g Lipinska et al. 2015; Nozaki 1996). There are also many examples within specific groups in which discrete differences in morphology and/or behaviour coexist within the same population. For instance, the switch from solitary to swarming behaviour in locusts (reviewed in Simpson et al. 1999), migratory and sedentary behaviours in the rainbow trout (Zimmerman & Reeves 2000; Hecht et al. 2013) or different morphs of males in the ruff (Widemo 1998) amongst numerous other examples.

In many cases, a trait that could result in an adaptive advantage for one of the phenotypes, results in detrimental effects for other phenotypes in the population. For instance, in *Drosophila melanogaster*, males seek to mate as much as possible. Mating, however, has a fitness cost in females, mediated through proteins in the seminal fluid. These proteins increase egg production in females and destroy sperm from other males, but it decreases female lifespan (Chapman et al. 1995). Such cases of competing evolutionary pressures between sexual phenotypes are known as sexual conflict (Chapman et al. 2003). Sexual conflict is a particular case of evolutionary conflict, which can be extended to any cases where several discrete phenotypes coexist in a population. At the molecular level, evolutionary conflict also impacts the genome, because if alternative phenotypes exist in the same population, they often share most, if not all, of their genome. The group of processes that reflect evolutionary conflict in the genome are known as genomic conflict (Hurst 1992). Genomic conflict can result in instances where, for example, several alleles co-adapted to a specific phenotype spread separately due to recombination, resulting in potential allelic incompatibilities within the same genome. It has been hypothesised that instances of genomic conflict may have led in several cases to selection for low recombination between co-adapted loci, resulting in low recombination regions of the genome linking co-adapted alleles. These low recombination regions are known as supergenes, and they are linked to several cases of stable polymorphisms within a population (Thompson & Jiggins 2014). Theory predicts that once a low recombination region linked to a specific phenotype has emerged, selection will favour additional conflicting loci to become linked to such region, resulting in the growth of the supergene region, with several phenotype-specific alleles linked to it (Charlesworth 2016). This theoretical framework would explain, for instance, the enrichment of sexually biased loci in many sex chromosomes (Mank 2017). This idea has

been challenged more recently in light of the evidence supporting the expansion of supergene regions without apparent addition of conflicting loci. Supergene expansion would, instead, depend on processes emerging from the lack of recombination (Branco et al. 2018; Dufresnes et al. 2015). These ideas are not, however, mutually exclusive and, in fact, are likely to be acting in conjunction (as the results from the previous chapter suggest). In either case, evolutionary conflict seems to be a key factor in supergene evolution.

A particular case where evolutionary conflict is likely to arise is that of local adaptation with gene flow. These are scenarios where several populations are locally adapted to their specific environment, but they are all connected by gene flow. As a consequence, alleles which are adapted to a specific ecological and genetic environment can enter a population where they would be maladaptive. Such a scenario would favour a linked group of genes with a set of alleles co-adapted to a specific environment (Kirkpatrick & Barton 2006). This type of selection is thought to be the evolutionary process taking place in Atlantic cod (*Gadus morhua*) populations. This species lives along the Atlantic coast of North America, a broad range that includes many diverse ecotypes. There is intense gene flow between their populations, and yet each population needs specific physiological and behavioural adaptations for their particular ecotype. A large proportion of the total genetic divergence between populations are localised in diversity islands of tightly linked genes within a number of inversions (Barney et al. 2017; Hemmer-Hansen et al. 2013). Additionally, many of the genes associated with local adaptations are localised in these inversions (Clucas et al. 2019) in what could be considered as a putative supergene.

Both antagonism between discrete phenotypes in a population and gene flow between locally adapted populations, are thus likely to lead to evolutionary conflict. Evolutionary conflict can, in turn, lead to the formation of supergenes under specific circumstances. More precisely, the emergence of supergenes would be limited to cases where the antagonistic loci are initially under some form of prior linkage. For instance, genes which are in close physical proximity in the genome (Charlesworth & Charlesworth 1975). It is important to note that antagonism between phenotypes and gene flow in the context of local adaptation are not necessarily mutually exclusive processes, and could (and almost certainly do) act at the same time, under different conditions. The relevant question then is not which of these processes underlies the spread of antagonistic alleles, but rather under which circumstances either of these processes plays a relatively more important role.

One way of answering this question is to focus on genes under intralocus evolutionary conflict. That is, genes with alleles having different fitness effects depending on the phenotype in which they are expressed (for a review with examples on sexual conflict, see

(Bonduriansky & Chenoweth 2009). Genes under intralocus evolutionary conflict are easier to study than entire antagonistic traits. This is because whole traits, tend to be encoded by several genes which are likely to interact with each other. These interactions make it more difficult to predict how allele frequencies may change through time and vary between different phenotypes. The frequencies of different alleles for a single locus, however, are easier to predict and to measure. Carrying out experiments to assess the relative roles of different evolutionary processes in the spread of antagonistic alleles is hard, especially because specific antagonistic alleles are rarely identified outside model organisms such as *D. melanogaster* (e.g. Innocenti & Morrow 2010). Evolution would need to be measured in real time in several populations under different conditions of gene flow and selection. Instead, statistical models using parameters measured in natural populations can be used as tools to simulate different evolutionary scenarios and their impacts in allele frequencies.

Such models have been used before to explore the spread of antagonistic alleles in sex determination systems. For instance, Veltsos et al. (2008) used an analytical approach to determine how and when sexually antagonistic alleles may spread in a population with and without a new sex chromosome. Models of sex determination describe a system in which gene flow between phenotypes is constant. The role of different types of gene flow between phenotypes for the spread of antagonistic alleles remains understudied.

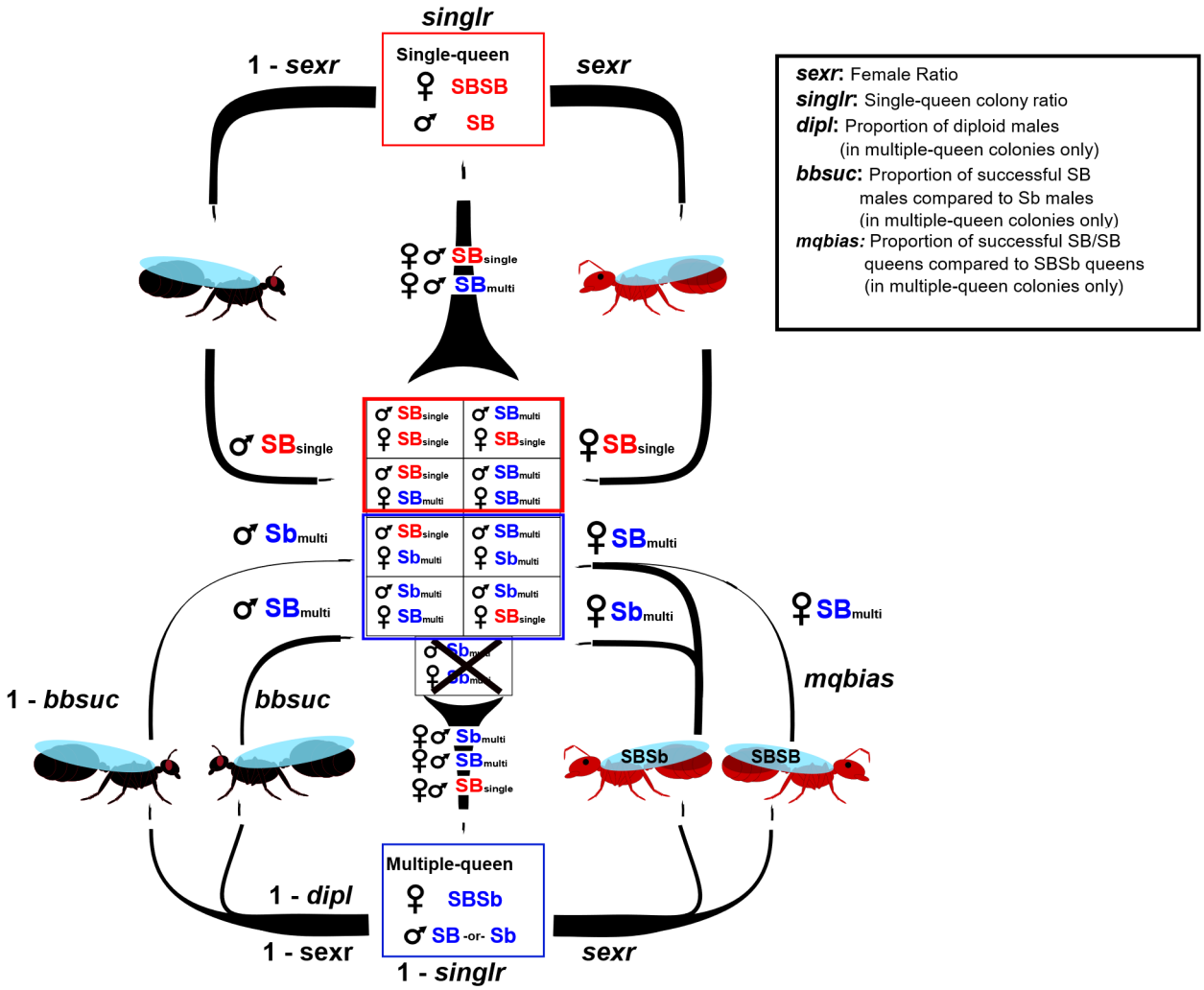


Figure 3.1: Schematic representation of the gene flow between social forms. Single queen colonies produce two types of gametes, male gametes carrying the SB variant (♂SB_{single}) and female gametes carrying the SB variant (♀SB_{single}). Multiple-queen colonies produce four types of gametes, male gametes carrying the SB (♂SB_{multi}) or Sb (♂Sb_{multi}) variant and females carrying the SB (♀SB_{multi}) or Sb variant (♀Sb_{multi}). The thickness of the arrows represents the relative proportions of each of those gametes in the population. That proportion is controlled in the model by the parameters *mqbias*, *bbsuc*, *singlr* and *sexr*, all defined in the box below.

Here I model the spread of antagonistic alleles in a population of the fire ant *Solenopsis invicta* with two discrete phenotypes linked to a supergene and under different regimes of gene flow. *S. invicta* is a good model to study the relative effects of gene flow and evolutionary conflict in the spread of antagonistic alleles. Throughout this chapter, as described in the introduction to this thesis, I will consider the colony as the level of selection. Based on the idea of kin selection (Hamilton, 1964a, Hamilton, 1964b), because all individuals of the same colony will be highly related, selection should act in the same way throughout the colony. As a result, a single colony can be considered as a selection unit. This species has two colony-level phenotypes, multiple and single-queen colonies, linked to a supergene with two variants SB and Sb. This system can therefore be explored to study the effects of antagonistic alleles with and without linkage to a supergene haplotype. In addition, the gene flow between social forms is not necessarily constant, nor unidirectional (Fig 3.1). That is, depending on several factors such as the differences in the fitness of males carrying either variant of the supergene or the culling of specific genotypes (reviewed in Chapter 1), gene flow between colony types can vary substantially.

The model described here will quantify how an antagonistic allele spreads as a function of the magnitude of the intensity of selection and gene flow. Because antagonistic alleles can exist with and without linkage to a supergene, the model also tests the effect of linkage under different gene flow scenarios.

The parameters used to characterise the modelled populations of *S. invicta* are taken from empirical data extracted from natural populations. The results of these simulations show that gene flow has a strong impact on the outcome of the spread of antagonistic alleles. We suggest that selection for linkage of antagonistic loci may have emerged as a mechanism to buffer the impact of variable gene flows between social forms.

Methods

The model simulates the changes in allele frequency at a single locus with two alleles in a population of *S. invicta* with both single and multiple-queen colonies. The alleles of the focal locus have fitness effects that differ between single- and multiple-queen colonies. Simulations were run for a locus linked to the supergene, and repeated for an unlinked locus. Because the supergene determines the phenotype of the colony (single or multiple queen), linkage to the supergene modifies the proportion of time that each allele spends in each social form.
The model is not spatially explicit and is deterministic, hence neglecting the effects of genetic drift.

Modelling allele frequency changes in *S. invicta*

The alleles at the focal locus were designated ‘c’ and ‘d’. The model is represented in Fig 3.2. It assumes that a first round of selection acts at the level of the colony. This selection causes changes in allele frequencies calculated using the standard model for selection in diploids (Hartl et al. 2007) reflecting the genotype of the queen. Because the focal locus has two alleles, there are three possible genotypes: homozygotes ‘cc’ and ‘dd’ and heterozygotes ‘cd’. Each genotype has a fitness effect (ω) associated with it. The average fitness (ϖ) of the population at a given time *t* depends on the frequencies of each allele (*p* for the frequency of allele ‘c’ and *q* for allele ‘d’, where $p = 1 - q$) and on the fitness effects of each genotype, such that:

(I) $\varpi = p^2\omega_{cc} + 2pq\omega_{cd} + q^2\omega_{dd}$

The change in frequency of allele ‘c’ in time *t*+1 (*p’*) depends on the fitness effect of that particular allele relative to the average fitness, such that after selection the allele frequency is:

(II) $p' = \frac{p^2\omega_{cc} + pq\omega_{cd}}{\varpi}$

The fitness parameters in (I) and (II) depend on the colony phenotype (single or multiple-queen). The relative proportion of each colony phenotype in the population is fixed and controlled by the parameter *singlr*, that is, the proportion of single-queen colonies in the

population. The total changes in allele frequencies in the population are calculated based on the phenotype-specific ones and the relative proportion of each social form such that:

(III) $p' = singlr \times p'_{sing} + (1 - singlr) \times p'_{multi}$

The model also takes into account the potentially asymmetric gene flow between social forms. The contribution of each colony type and the allele frequencies are changed by a second round of selection affecting the multiple queen colonies, determined by the parameters *sexr, dipl, bbsuc* and *mqbias.*

Selection on males

The fitness of multiple-queen colonies is reduced by the occurrence of diploid males (Tschinkel 2006). These males do not reproduce and therefore do not contribute to the gene pool of the next generation. The parameter *dipl* represents the proportion of diploid males in the multiple-queen populations, so the effective production of males is (1-*dipl*)(1-*sexr*); *sexr* being the (female) sex ratio in single-queen colonies. In addition, two effects modify the relative contribution of Sb males: they are less efficient at fertilising queens but, on the other hand, SB males may be culled before emergence (Fritz et al. 2006) The combined effect is quantified by a parameter giving the proportion of fertilizations by Sb males, *bbsul*.

Selection on queens

Selection in multiple-queen colonies also acts on the production of new queens. A large proportion contain the the Sb allele (DeHeer 2002). At least two factors contribute to this bias: they are likely to be killed if they are homozygous for the SB variant of the supergene (Keller & Ross 1998); additionally, eggs carrying the Sb allele are more likely to become queens (Buechel et al. 2014). The parameter *mqbias* ∈ [0,1] quantifies this effect: if *mqbias* = 0, multiple-colonies produce only SB/Sb queens; if *mqbias* = 1, they produce a 50-50 proportion of each queen genotype.

Recurrence equations

The core of the model comprises of five variables recording the relative frequency of the key types of gamete. These frequencies are written p_x where the subscript x descriminates between single (*sing*) and multiple-queen (*multi*) colonies, sex, and in the case of female gametes, the Sb/Sb genotype: *multi*♂, *multiSb*♀, *multiSB*♀ *sing*♂ and *sing*♀.

Only queens which are homozygote SB for the supergene and males carrying the SB variant contribute to the next generation of single-queen colonies. The first step in obtaining the frequency of the *c* allele after selection is to calculate the mean fitness based on equation (I):

(IV)

$$cc_{sing} = p_{sing}p_{sing} + bbsuc \times p_{multi}p_{sing} + bbsuc \times p_{multi}p_{multiSB} + p_{sing}p_{multiSB}$$

$$cd_{sing} = p_{sing}q_{sing} + bbsuc \times p_{multi}q_{sing} + bbsuc \times p_{multi}q_{multiSB} + q_{sing}p_{multiSB} + q_{sing}p_{sing} + bbsuc \times q_{multi}p_{sing} + bbsuc \times q_{multi}p_{multiSB} + p_{sing}q_{multiSB}$$

$$dd_{sing} = q_{sing}q_{sing} + bbsuc \times q_{multi}q_{sing} + bbsuc \times q_{multi}q_{multiSB} + q_{sing}q_{multiSB}$$

$$\varpi_{sing} = cc_{sing} + cd_{sing} + dd_{sing}$$

The change in allele frequency for allele 'c' in the next generation for single queen colonies (p'_{sing}) is then calculated similarly to (II):

(V) $$p_{sing}' = \frac{cc_{sing}\omega_{ccsing} + 0.5 \times cd_{sing}\omega_{cdsing}}{\varpi_{sing}}$$

In multiple-queen colonies, queens carrying the Sb allele would produce a multiple-queen offspring regardless of the genotype of the male they would mate with, because males are haploid. Their offspring would, therefore, always be heterozygote SB/Sb or homozygote Sb/Sb. The only SB/SB queens that could form multiple-queen colonies would be those mating with Sb males. Using the same parameters as in IV to account for the relative proportion of Sb males and SB/SB queens from multiple-queen colonies, the changes in allele frequencies in multiple-queen colonies would therefore be calculated as follows:

(VI)

$$cc_{multi} = p_{sing}p_{multiSb} + p_{multi}p_{multiSb} + (1 - bbsuc) \times (p_{multi}p_{multiSB} + p_{multi}p_{sing})$$

$$cd_{multi} = p_{sing}q_{multiSb} + p_{multi}q_{multiSb} + (1 - bbsuc) \times (p_{multi}q_{multiSB} + p_{multi}q_{sing}) + q_{sing}p_{multiSb} + q_{multi}p_{multiSb} + (1 - bbsuc) \times (q_{multi}p_{multiSB} + q_{multi}p_{sing})$$

$$dd_{multi} = q_{sing}q_{multiSb} + q_{multi}q_{multiSb} + (1 - bbsuc) \times (q_{multi}q_{multiSB} + q_{multi}q_{sing})$$

$$\varpi_{multi} = cc_{multi} + cd_{multi} + dd_{multi}$$

(VII)

$$p_{multi}{}' = \frac{cc_{multi}\omega_{ccmulti} + 0.5 \times cd_{multi}\omega_{cdmulti}}{\varpi_{multi}}$$

Note that, for the sake of simplicity, this model does not account for the lethality of Sb/Sb queens.

Once the change in frequencies has been calculated in both phenotypes, the overall allele frequencies in the population is calculated using (III). The equations above do not account for linkage. Linkage was added following the method used in a similar method in (Veltsos et al. 2008), where the authors modelled the spread of an antagonistic allele linked to sex chromosomes in a population with sexual dimorphism. Linkage is added by partitioning the frequencies of the alleles in the focal locus between phenotypes and variants. In this case, the frequencies of alleles 'c' and 'd' need to be partitioned further between Sb (p_{Sb} for allele 'c' and q_{Sb} for allele 'd') and SB in multiple ($p_{multiSB}$ for allele 'c' and $q_{multiSB}$ for allele 'd') and single (p_{singSB} for allele 'c' and q_{singSB} for allele 'd') queen colonies.

In this model, the frequencies for each gamete are calculated taking into account the same parameters described above affecting the gene flow between phenotypes. In this case, however, 6 instead of 4 gametes need to be explicitly modelled per allele: SB in males and queens from either multiple or single-queen colonies and Sb in males and queens from multiple-queen colonies, represented in the model as *SBmulti*♂, *SBmulti*♀, *SBsing*♂, *SBsing*♀, *Sb*♂and *Sb*♀ respectively. Changes in allele frequencies for each partition in *t+1* are then calculated using variations of (II) to account for the fact that each new colony needs to be formed by a queen and a male gamete from the right genotype. The average fitness is calculated independently for multiple (ϖ_{multi}) and single-queen colonies (ϖ_{sing}):

(VIII)

$$cc_{multi} = p_{Sb}\,p_{SBsing}\,♂ + p_{SBmulti}\,♂p_{Sb}\,♀ + p_{SBmulti}\,♀p_{Sb}\,♂ + p_{SBsing}\,♀p_{Sb}\,♂$$

$$cd_{multi} = p_{Sb}\,♀q_{SBsing}\,♂ + p_{SBmulti}\,♂q_{Sb}\,♀ + p_{SBmulti}\,♀q_{Sb}\,♂ + p_{SBsing}\,♀q_{Sb}\,♂ +$$
$$q_{Sb}\,♀p_{SBsing}\,♂ + q_{SBmulti}\,♂p_{Sb}\,♀ + q_{SBmulti}\,♀p_{Sb}\,♂ + q_{SBsing}\,♀p_{Sb}\,♂$$

$$dd_{multi} = q_{Sb}\,♀q_{SBsing}\,♂ + q_{SBmulti}\,♂q_{Sb}\,♀ + q_{SBmulti}\,♀q_{Sb}\,♂ + q_{SBsing}\,♀q_{Sb}\,♂$$

$$\varpi_{multi} = cc_{multi} + cd_{multi} + dd_{multi}$$

(IX)

$$cc_{sing} = p_{SBsing}p_{SBsing} + p_{SBsing}p_{SBmulti} + mqbias \times (p_{SBmulti}p_{SBsing} + p_{SBmulti}p_{SBmulti})$$
$$cd_{sing} = p_{SBsing}q_{SBsing} + p_{SBsing}q_{SBmulti} + mqbias \times (p_{SBmulti}q_{SBsing} + p_{SBmulti}q_{SBmulti}) + q_{SBsing}p_{SBsing} + q_{SBsing}p_{SBmulti} + mqbias \times (q_{SBmulti}p_{SBsing} + q_{SBmulti}p_{SBmulti})$$
$$dd_{sing} = q_{SBsing}q_{SBsing} + q_{SBsing}q_{SBmulti} + mqbias \times (q_{SBmulti}q_{SBsing} + q_{SBmulti}q_{SBmulti})$$

$$\varpi_{sing} = cc_{sing} + cd_{sing} + dd_{sing}$$

Note that, in this case, SbSb individuals are not considered, as they are assumed in the model to have no impact over evolutionary time. The change in 'c' allele frequencies for alleles linked to SB in single-queen colonies (p'_{SBsing}) is then calculated as in (V). For multiple-queen colonies, on the other hand, two changes in allele frequencies 'c' need to be calculated, one for Sb (p'_{Sb}) and another for SB ($p'_{SBmulti}$), both using ϖ_{multi}. To calculate these parameters, only the variant with allele 'c' has to be considered, so that:

(X)

$$p'_{SBmulti} = \frac{cc_{multi}\omega_{ccmulti} + (q_{Sb}p_{SBsing} + p_{SBmulti}q_{Sb} + p_{SBmulti}q_{Sb} + p_{SBsing}q_{Sb})\omega_{cdmulti}}{\varpi_{multi}}$$

$$p_{Sb}' = \frac{cc_{multi}\omega_{ccmulti} + (p_{Sb}q_{SBsing} + q_{SBmulti}p_{Sb} + q_{SBmulti}p_{Sb} + p_{SBsing}q_{Sb})\omega_{cdmulti}}{\varpi_{multi}}$$

Once each partition of the changes in frequency of allele 'c' are calculated, they are added together by calculating first the overall frequency in SB ($p'SB$), using the proportions of each phenotype in the population. Single-queen colonies have two copies of SB, and multiple-queen colonies only one, so to calculate the overall frequency in SB:

(XI) $$p'_{SB} = \frac{singlr \times 2}{singlr \times 2 + (1 - singlr)} \times p'_{SBsing} + \frac{(1 - singlr)}{singlr \times 2 + (1 - singlr)} \times p'_{SBmulti}$$

Similarly, the calculation of the overall change in allele frequency in the population (p') needs to take into account the proportions of each colony. Additionally, SB is present twice in single-queen colonies, but SB and Sb only once in multiple-queen colonies, so that:

(XII) $$p' = \frac{2 \times singlr + (1 - singlr)}{2} \times p'_{SB} + \frac{(1 - singlr)}{2} \times p'Sb$$

It is important to note that this model does not simulate the evolution of linkage. Instead, linkage is assumed to be already present in the population, and fully associated to social form (for the theoretical framework regarding the circumstances of linkage evolution in the context of evolutionary conflict: (Charlesworth & Charlesworth 1975; Charlesworth & Charlesworth 1979; Turner 1967).

Simulations and input values

The models described above were run with a set of specific parameters to explore different scenarios of gene flow with and without linkage. The parameters *sexr* and *singlr* were left constant at 0.5. That is, in all simulations the assumption is that there are the same proportion of queens and males and the same proportion of single and multiple-queen colonies in the population. The former assumption is not necessarily true at the colony level (e.g. Passera et al. 2001), but across a whole population of *S. invicta* and over evolutionary time it is safer to assume a balanced sex ratio. The assumption of a balanced proportion of each social form requires more justification. If left undisturbed, multiple-queen colonies tend to over-perform single-queen colonies. As a result, multiple-queen colonies dominate in many populations of *S. invicta*, especially in the invasive North American range (Porter 1993; Glancey et al. 1987). Single-queen colonies are better at colonising new habitats, were other colonies have not yet been settled, or previous colonies have been removed by disturbance (DeHeer 2002). This means that there might be a constant turnover of social forms in a cycle of establishment of single-queen colonies, gradual replacement by multiple-queen colonies, disturbance, and, again, establishment of new single-queen colonies. Such turnover has been observed in invasive populations of North America, albeit at a relatively slow rate (Porter 1993). In the native South American range, several populations display a varying proportion of either social form (Ahrens et al. 2005). The model thus assume that, over evolutionary time, the average proportion of either social form should be balanced.

The proportion of diploid males in multiple-queen colonies (*dipl*) has been shown to be very high in invasive populations (near 100% (Ross & Shoemaker 1993). This is, however, an artifact of the low genetic diversity of the invasive populations of *S. invicta* (Ross et al. 1993). In the native South American populations, where most of *S. invicta* evolution has taken place the proportion of diploid males is estimated to be around 14% in multiple-queen populations (Ross et al. 1993). The model will therefore assume this value as a constant over evolutionary time (*i.e. dipl* = 0.14).

There will be two scenarios of gene flow tested here:

- High gene flow between social forms (no bias of SB sexuals in multiple-queen colonies)
- Low gene flow between social forms (bias of SB sexuals in multiple-queen colonies)

The first scenario will assume that all SB males and SB/SB queens from multiple-queen colonies survive and are able to mate. Sb males have been shown to perform worse than their SB counterparts in multiple-queen colonies. Sb males have a lower sperm count than SB males. All else being equal, SB males are estimated to sire 71% of the offspring (Lawson et al. 2012). Under this scenario, *mqbias* is set to 1 (there is a 50-50 SB/SB-SB/Sb split in queens coming from multiple-queen colonies) and *bbsuc* to 0.71 (71% of SBSB males are successful at producing offspring). This scenario is not realistic given the existing evidence on *S. invicta*. Istead, it is used as an extreme model where gene flow between social forms is at its maximum.

For the second scenario, the model assumes the bias of sexuals that do not carry the Sb allele in multiple-queen colonies. According to (Keller & Ross 1998), 61% of SB/SB queens aged between 7-10 days were killed in multiple-queen colonies. Additionally, according to data from (Buechel et al. 2014), only 24% of SB/SB larvae develop into queens, out of the expected 50%, that is a 48% bias against queens not carrying the Sb allele. Therefore, in the model, out of the total potential proportion of SB/SB queens in multiple queen colonies, only 48% will develop, and from those only 39% will survive, resulting in a total 18.72% of bias against queens not carrying the Sb allele. For males, this proportion is estimated to be around 93% based on data from (Fritz et al. 2006). In the model, *mqbias* is thus set to 0.19 (rounded from 0.1872) and *bbsuc* to 0.07.These numbers are likely to be variable, given that the aggressivity towards sexuals not carrying the Sb allele may be linked to the proportion of SBSb workers in the colony (Fritz et al. 2006). Nonetheless, these estimates represent a first approximation for the modelling of the complex and dynamic gene flow of *S. invicta* based on empirical data. The two scenarios were run with and without linkage.

Finally, each simulation models different levels of fitness effects for all genotypes in either social form. In all simulations, the assumption is that the ‘c’ allele is beneficial for single-queen colonies and potentially detrimental for multiple-queen colonies and vice-versa for the ‘d’ allele. The fitness for each genotype within social form is relative to the rest of the population. Relative fitness is calculated for each genotype within social form using two

additional parameters (Hartl et al. 2007): heterozygous effect (*h* common to the whole population) and the selection coefficient (s_{multi} and s_{singl}, specific to each social form). The relative fitnesses in each social form are calculated as follows:

(XIII)

$$\omega_{ccsing} = 1;\ \omega_{cdsing} = 1 - s_{sing}h;\ \omega_{ddsing} = 1 - s_{sing}$$
$$\omega_{ccmulti} = 1 - s_{multi};\ \omega_{cdmulti} = 1 - s_{multi}h;\ \omega_{ddmulti} = 1$$

The parameter *s* is therefore the relative fitness effect of the detrimental allele in each social form. It can vary between 0 and 1, meaning that the homozygote for the detrimental allele has either a neutral or a detrimental effect on the fitness on the bearer. The parameter *h* defines the effect of the detrimental allele in heterozygotes. It varies between 0 and 1, if set to 0, the detrimental allele is a complete recessive, if set to 1, a complete dominant. If *h* is set to 0.5, both alleles are strictly additive.

The simulations ran for all the combinations of selection coefficients values in each social form between 0 and 0.1, with step increases of 0.001. That is, the simulations tested for scenarios were selection goes from no effect, to an effect of 10% on fitness. The heterozygous effect *h* was kept at 0.5 for all simulations. All simulations ran for 10.000 generations, time enough for allele frequencies to reach an equilibrium, confirmed by running several prior test simulations. The results report the final frequency of the allele 'c' after 10.000 generations for all combinations tested of s_{multi} and s_{sing}. The initial allele frequency for 'c' (*p*) was set at 0.1 for all simulations.

All the simulations were built and ran in R (v3.6.1; R Core Team, 2019), using the package "plotly" (v4.9.0; Sievert 2018) for visualisation of the results.

Results

High gene flow, no linkage scenario

Multiple simulations were run under the scenario of high gene flow between social forms, that is, without bias for Sb individuals in multiple-queen colonies, for 10,000 generations and different combinations of selective pressures in single and multiple-queen colonies.
When the focal antagonistic locus is not linked to the supergene (Fig 3.2), its alternative alleles are rarely kept at an equilibrium. Allele 'c' is either fixed or lost in the population for

most combinations of selective pressures. The range of selection pressures at which both alleles exist in the population becomes wider as selection pressures increase. Some sort of equilibrium where both ‘c’ and ‘d’ alleles are present in the population is reached only when selection against the ‘d’ allele in the single-queen colony phenotype (s_{sing}) is between 1.2 and 1.5 times greater than selection against the ‘c’ phenotype in the single-queen colony phenotype (s_{multi}). Put differently, selection needs to be up to 50% stronger in single-queen colonies to match the effect of selection in multiple-queen colonies. Therefore, under the high gene flow scenario without linkage, multiple-queen colonies are at an advantage with respect to single-queen colonies: all else being equal, alleles beneficial for multiple-queen colonies are more likely to become fixed in the population.

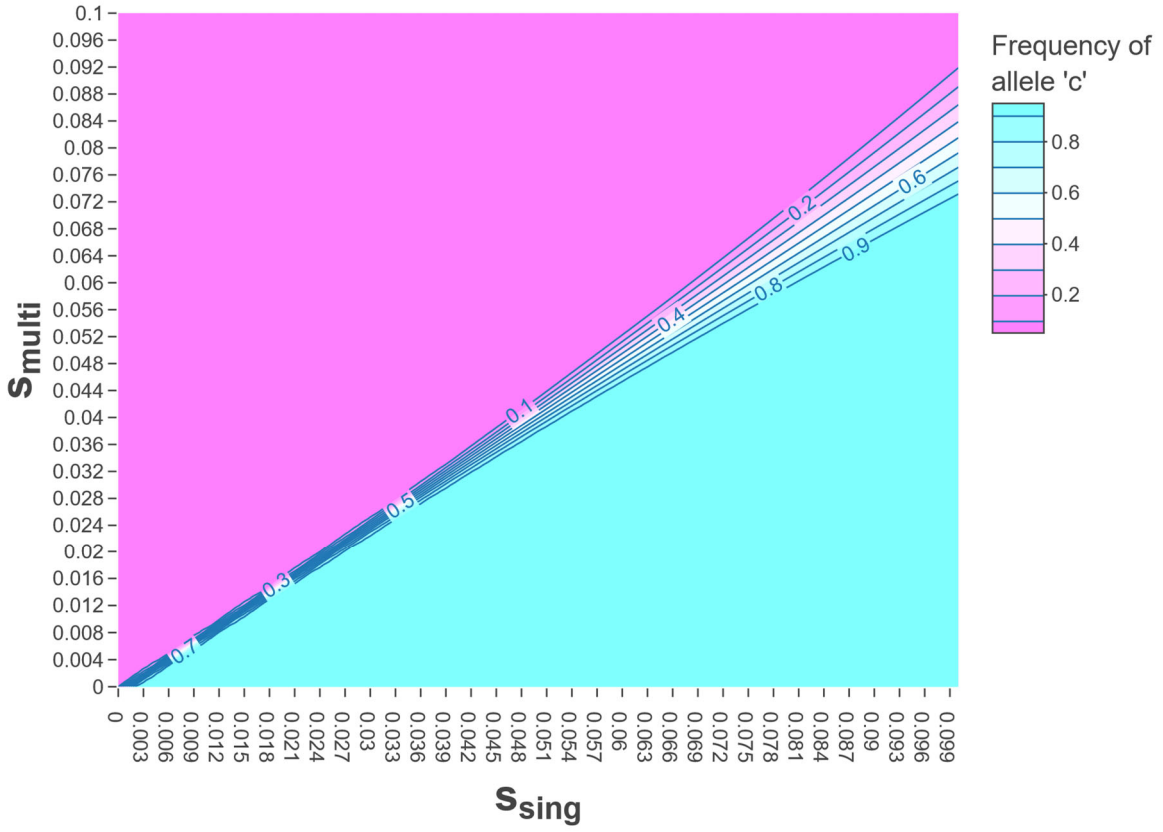


Figure 3.2: Antagonistic allele frequencies of the allele ‘c’ after 10,000 generations under high gene flow between social forms and without linkage of the antagonistic locus to the supergene. The simulations were run under different fitness effects of allele ‘c’ against multiple-queen colonies (s_{multi}) and of allele ‘d’ against single-queen colonies (s_{single}). Darker tones indicate either high (blue) or low (pink) frequencies of ‘c’. Lighter colours indicate combinations of fitness where both alleles coexist. The lines in the plot indicate combinations of fitnesses for which the allele frequencies are the same. A larger surface of dark pink combinations indicates that the ‘c’ allele is lost in most cases, that is, multiple-queen colonies tend to be at an advantage.

High gene flow, linkage to supergene scenario

Under the same high gene flow scenario, but with linkage of the antagonistic locus to the supergene (Fig 3.3a), the resulting allele frequencies seem to be inverted. The multiple-queen beneficial allele (‘d’) is less likely to become fixed in the population. There is, however, an important difference, the single-queen beneficial allele (‘c’) never becomes fixed in the population.

As soon as allele ‘d’ has a beneficial effect in multiple-queen colonies, it becomes fixed in the Sb variant of the supergene (Fig 3.3b), which is only present in multiple-queen colonies. Conversely, in the SB supergene variant (Fig 3.3c), allele ‘d’ becomes lost in most combinations of selection pressures. More specifically, selection pressure in multiple-queen colonies needs to be between 1.3 and 1.5 to reach an equilibrium where both alleles ‘c’ and ‘d’ are present in the SB variant population.

In all simulations the proportion of either social form was set to 50% and the SB variant is present twice in single-queen colonies and once in multiple-queen colonies, SB is present in 75% of the population, and Sb in 25%. Because ‘c’ is more likely to be fixed in SB and this variant makes out ¾ of the population, the single-queen beneficial allele is more likely to be found at high frequencies in the whole population. If the ‘c’ allele has any detrimental effect on multiple-queen colonies, however, it never becomes fixed in the whole population, because it would be lost in Sb, which makes out ¼ of the population. Consequently, the ‘c’ allele can reach at most a frequency of 0.75 in the population. This scenario is, therefore, slightly beneficial for single-queen colonies, because its beneficial allele tends to be more common, but on the other hand, the multiple-queen beneficial allele is never lost.

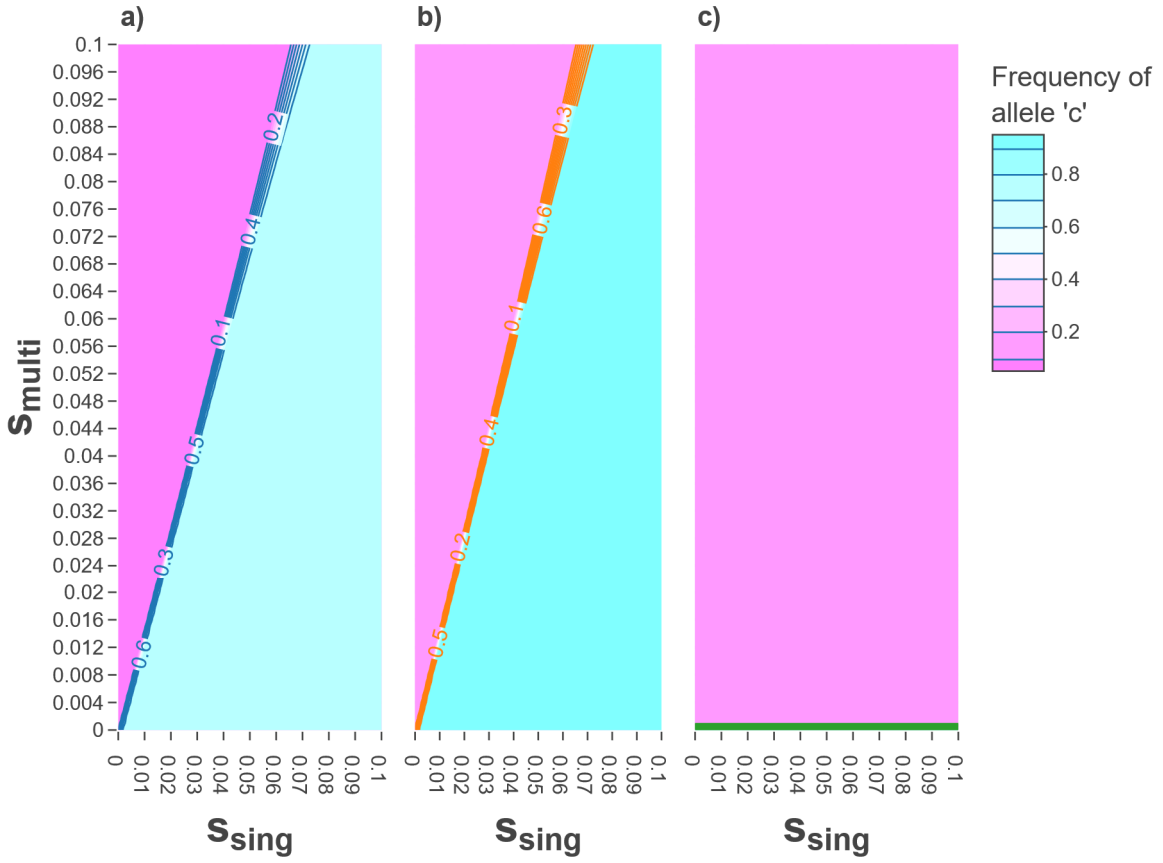


Figure 3.3: Antagonistic allele frequencies of the allele ‘c’ after 10,000 generations under high gene flow between social forms with the antagonistic locus being linked to the supergene. The figure shows the results for the overall population (**a**), the SB variant of the supergene (**b**) or the Sb variant of the supergene (**c**). The simulations were run under different fitness effects of allele ‘c’ against multiple-queen colonies (s_{multi}) and of allele ‘d’ against single-queen colonies (s_{single}). Darker tones indicate either high (blue) or low (pink) frequencies of ‘c’. Lighter colours indicate combinations of fitness where both alleles coexist. The lines in the plot indicate combinations of fitnesses for which the allele frequencies are the same. There is a larger surface of blue combinations in (**a**), indicating that the ‘c’ allele is

Low gene flow, no linkage scenario

The resulting allele frequencies for the simulations with low genetic flow between social forms and without linkage (Fig 3.4), are almost opposite to those described in the high gene flow scenario. Again, there are very few combinations of selection pressures where both alleles of the antagonistic locus coexist. In this case however, the single-queen beneficial allele ‘c’ becomes fixed in most of the cases. Selection needs to be around 3 times higher in multiple-queen colonies compared to that in single-queen colonies to maintain both alleles in the population. Below this threshold, the multiple-queen beneficial allele ‘d’ is always lost. This scenario is thus highly beneficial for single-queen colonies.

Low gene flow, linkage to supergene scenario

Under the same scenario but with the antagonistic locus being linked to the supergene (Fig 3.5a), the results are not very different from those reported in the high gene flow scenario. Again, the ‘c’ allele is more likely to be found at high frequencies in most combinations of selection, but never reaches frequencies higher than 0.75, because it is lost in Sb (Fig 3.5b). The main difference with respect to the high gene flow scenario is that the single-queen beneficial allele reaches high frequencies in more cases. Under the low gene flow scenario, selection pressure against the ‘c’ allele in multiple-queen colonies needs to be at least twice the selection pressure against the ‘d’ allele in order for both alleles to co-exist in the SB population (Fig 3.5c). Overall, the low gene flow scenario with linkage is beneficial to single-queen colonies, and slightly more so than the high gene flow scenario. This difference, however, is much larger between the two scenarios without linkage.

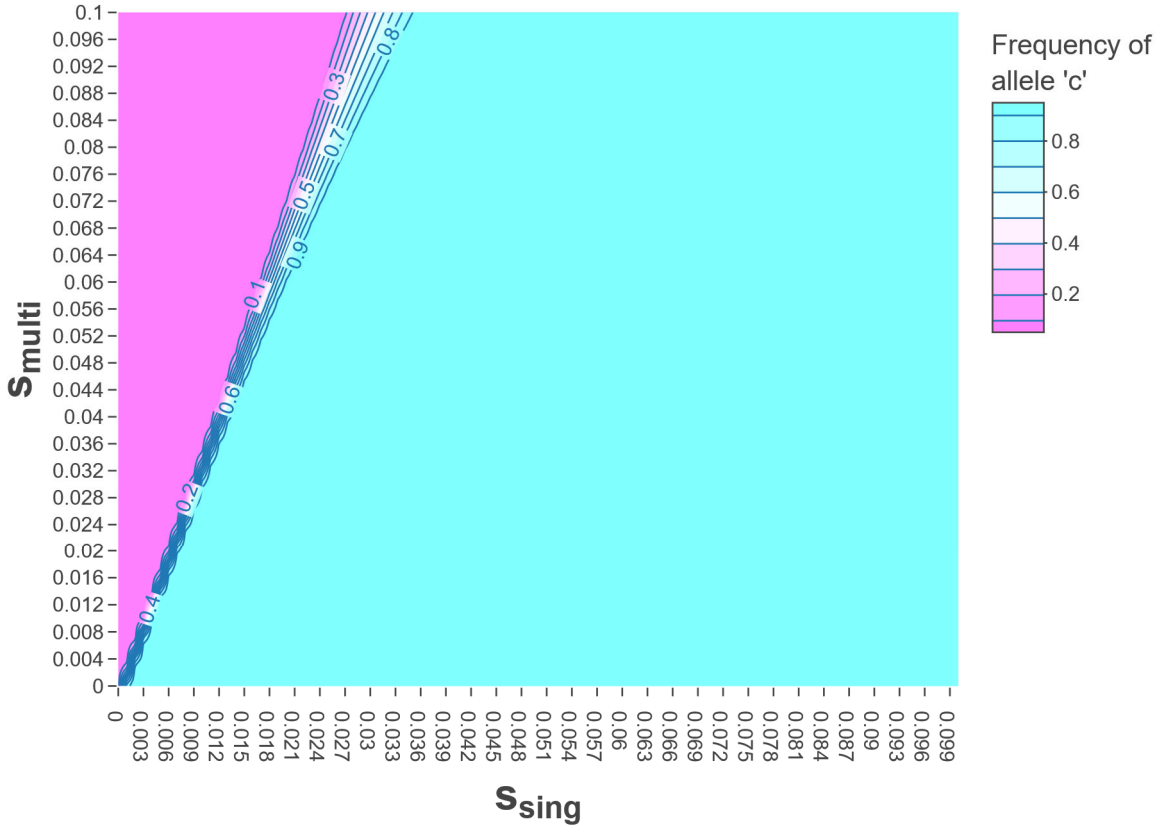


Figure 3.4: Antagonistic allele frequencies of the allele ‘c’ after 10,000 generations under low gene flow between social forms and without linkage of the antagonistic locus to the supergene. The simulations were run under different fitness effects of allele ‘c’ against multiple-queen colonies (s_{multi}) and of allele ‘d’ against single-queen colonies (s_{single}). Darker tones indicate either high (blue) or low (pink) frequencies of ‘c’. Lighter colours indicate combinations of fitness where both alleles coexist. The lines in the plot indicate combinations of fitnesses for which the allele frequencies are the same.

Discussion

The balance between Sb and non-Sb carrying sexuals explains the impact of gene flow

These results can be understood as a consequence of the asymmetries in gene flow between the two colony types. First consider SB/SB queens and SB males. They can originate from both types of colony. All else being equal, however, a smaller proportion will be produced by the multiple-queen colonies, because they also produce heterozygous females and Sb males. Furthermore there is additional selection in multiple-queen colonies that reduces their production of SB queens and fertile males. Consequently, the single-

queen beneficial allele would be exposed to positive selection more often in this part of the population, and thus, would increase in frequency unless there were a stronger counter-selection during its residence in multi-queen colonies.

The second component of gene flow that needs to be considered is through gametes carrying the Sb allele. These are only produced by multiple-queen colonies (and are passed on to multiple queen colonies). Therefore, Sb gametes recycle through multiple-queen colonies each generation tending to favour the multiple-queen beneficial allele. The SB contribution, however, which makes up half the genome of multiple queen colonies, is a mixture from the two types of colony.

Without physical linkage between Sb and the antagonistic locus, in a scenario where most SB individuals are produced by single-queen colonies the equilibrium between antagonistic alleles would be tipped towards the single-queen beneficial variant. On the other hand, with greater gene flow, implying a higher contribution of multiple-queen colonies to SB gametes, the multiple-queen variant could be maintained with a smaller selective advantage.

Consequently, any factor affecting the relative strength of these two gene flows (either SB/SB queens and SB males or Sb carrying males and females) is likely to affect the outcome of selection on the antagonistic locus. Outcome, that for most values of selection involve either loss or fixation of an allele which is detrimental for an important part of the population (Figs 3.2 and 3.4). This pattern of quick loss/fixation of unlinked antagonistic alleles has been predicted in previous models (Charlesworth & Charlesworth 1975), and explains why supergenes seem more likely to arise in loci already under linkage. Specific combinations of any parameter affecting gene flow such as, the proportion of diploid males in multiple-queen colonies, bias towards the production of SBSb queens or the proportion of each colony type in the population will result in very different outcomes.

One the other hand, when the antagonistic locus is linked to the supergene that controls social form in *S. invicta*, gene flow has less of an impact on the outcome of selection. This resilience to gene flow occurs because in the model Sb is only present in multiple-queen populations. The antagonistic haplotype on Sb is, therefore, never exposed to selection in single-queen colonies. Put differently, influx of SB gametes by gene flow from single-queen colonies does not affect the antagonistic locus on the Sb chromosome. As a result, the multiple-queen allele always increases in frequency in Sb. The dynamics in the SB population are therefore similar to those described in the scenario without linkage, there is greater SB gene flow from the multiple-queen colonies. If the proportion of each social form is balanced at 50-50, SB will make out 75% of the population, and Sb, 25%, hence resulting

in high frequencies of the single-queen allele in most cases, but where both alleles co-exist in the population, regardless of the gene flow.

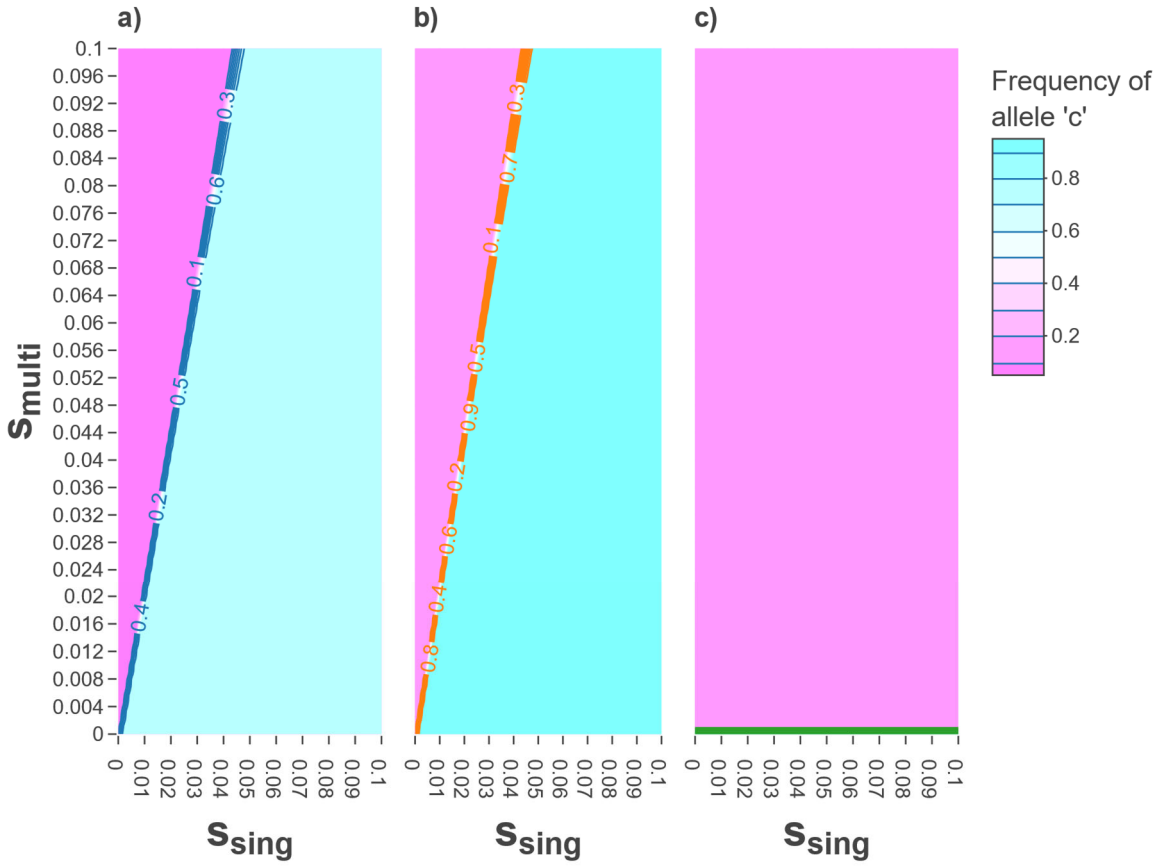


Figure 3.5: Antagonistic allele frequencies of the allele ‘c’ after 10,000 generations under low gene flow between social forms with the antagonistic locus being linked to the supergene. The figure shows the results for the overall population (**a**), the SB variant of the supergene (**b**) or the Sb variant of the supergene (**c**). The simulations were run under different fitness effects of allele ‘c’ against multiple-queen colonies (s_{multi}) and of allele ‘d’ against single-queen colonies (s_{single}). Darker tones indicate either high (blue) or low (pink) frequencies of ‘c’. Lighter colours indicate combinations of fitness where both alleles coexist. The lines in the plot indicate combinations of fitnesses for which the allele frequencies are

Gene flow as a driver of the evolution of the supergene

The results from the model show that allele frequencies at equilibrium are strongly affected by the gene flow between social forms of *S. invicta*. This gene flow is in turn, affected by several factors that are likely to vary over time. Here, we have tested two extreme scenarios, but in real populations of *S. invicta* many more intermediate scenarios are likely to be happening at the same time in different populations. For instance, the proportion of bias against sexuals not carrying the Sb variant in multiple-queen colonies is likely linked to the proportion of workers carrying such variant (Fritz et al. 2006; Buechel et al. 2014), and therefore variable from colony-to-colony. Other factors include the possibility of assortative

mating between individuals from the same colony type, suggested by some authors in *S. invicta* (Saddoris et al. 2016), and known to happen in other socially polymorphic ant species (Avril et al. 2019). The number of diploid males produced by single-queen colonies would also affect the gene flow between social forms, and therefore, the outcome of antagonistic allele frequencies. Again, the proportion of diploid males can vary strongly, owing to differences in genetic diversity among locations (Ross et al. 1993). For instance, in North American populations almost all males produced by multiple-queen colonies are diploid (Tschinkel 2006), resulting in an almost unidirectional gene-flow from single to multiple-gene colonies (Ross & Shoemaker 1993). Gene flow is not only likely to vary in space, but also in time. The proportion of the different social forms in a population is one of the parameters that is likely to affect the most gene flow. Owing to the life history of *S. invicta*, this proportion is likely to change over time. Initially, mostly single-queen colonies will colonise new environments. As population density increases, multiple-queen colonies outperform their single-queen counterparts and increase in frequency (Ross & Keller 1995). The proportion of each social form is highly variable in each population (Ross et al. 2007; Mescher et al. 2003), and changes over time (Porter 1993).

The model has shown that, without linkage to the supergene, there is a very narrow range of circumstances where both antagonistic alleles coexist in the population. In most situations one of the alleles becomes fixed. Additionally, given the high variability of gene flows between social forms, this specific set of circumstances is likely to be changing constantly over evolutionary time. Consequently, the most likely outcome for an antagonistic allele is to become lost or fixed. Evolutionary conflict for a specific locus is maintained in the population as long as two or more alleles with different fitness effects in different phenotypes coexist in the population (e.g. Chapman et al. 2003). The results discussed here are, therefore, consistent with the idea that evolutionary conflict on its own is unlikely to play a predominant role in the evolution of the social chromosome. Instead, the results of the simulations point at gene flow having a greater influence. To explore this statement in depth, we now need to focus on the results obtained in the simulations when the antagonistic alleles are linked to the supergene.

When linked to the supergene, allele frequencies in the antagonistic locus are more protected from gene flow. In all situations modelled, as long as an allele is beneficial to multiple-queen colonies, it became fixed in Sb. That is, the multiple-queen allele locus was never lost in the population, and all multiple-queen colonies had at least one copy of the beneficial allele. Linkage to the supergene, however, also comes at a cost for multiple-queen colonies. Because of SB and Sb do not recombine, all else being equal, the antagonistic

allele in the SB variant is twice as likely to be in a single-queen colony than in a multiple-queen colony. As a result, the single-queen allele tends to become fixed in SB, even when it has enough negative impact in multiple-queen colonies that it would be lost in the population under a no-linkage scenario. Because multiple-queen colonies have one copy of SB, in most cases all multiple-queen colonies carry an allele that may be detrimental for them. From a multiple-queen perspective, therefore, linkage to the supergene is beneficial, because it prevents the loss of multiple-queen beneficial alleles, but it comes at a cost because it makes it more likely to fix detrimental alleles in SB.

Here, we hypothesise that selection for the linkage of several antagonistic alleles stems from the effects that the supergene has in buffering the impacts of gene flow. According to this idea, selection for the *S. invicta* supergene originates as a way to “shield” beneficial alleles for multiple-queen colonies from the fluctuations in gene flow, even when fitness for these colonies could be higher without linkage at a specific point in time. For instance, consider a scenario where both antagonistic alleles coexist in the population. Let one of the alleles be slightly more beneficial for multiple-queen colonies, than it is detrimental for single-queen colonies. And finally, let there be, in the same population, colonies where the antagonistic locus is linked to a supergene, and colonies where it is not. At this point in time, the multiple-queen allele would be found at high frequencies in the single-queen colonies without linkage. In the colonies with linkage, it may be found at lower frequencies because it is lost in SB, but fixed in Sb. In this scenario, the colonies without linkage have a fitness advantage compared to those without. But now let there be a change in the gene flow between social forms. For instance, a flood that results in a reduction of multiple-queen colonies. The balance at the antagonistic locus is now tipped towards the single-queen allele. As a result, the multiple-queen allele is lost in populations without linkage, but it is still maintained in the populations with linkage. Now, the populations with linkage are at an advantage compared to the populations without linkage. If these changes in gene flow resulting in the loss of beneficial alleles occur often enough, linkage has a beneficial advantage over evolutionary time.

Alternatives to gene flow as a driver of supergene evolution: conflict and antagonism

Other than the effect of gene flow fluctuations, there are other explanations for the emergence of the supergene in *S. invicta* that are not mutually exclusive.

The previous section argued that evolutionary conflict was unlikely to play a dominant role in supergene evolution, because the fluctuations in gene flow would mean that antagonistic

alleles become either lost for most parameter combinations. It could be, however, that if the gene flow changes quickly enough, allele frequencies would not reach an equilibrium. In such a situation, evolutionary conflict could take place, and therefore selection for linkage of the antagonistic alleles, provided prior linkage (Charlesworth 2016). For conflict to arise in this way, the fluctuation in gene flow would need to be relatively frequent. The rate of replacement of single-queen colonies by multiple-queen colonies has been estimated at around 4%-6% per year in areas first colonised by single-queen colonies in the North American invasive range (Porter 1993). The proportion of each social form is therefore assumed to be relatively stable in short periods of time, which makes it unlikely that quick changes in gene flow could occur at a frequency high enough to maintain several antagonistic alleles in the population. It is important to note that this replacement estimation is based only on one study in the invasive range, where the life history of *S. invicta* may be slightly different to that of the native South American range (Ahrens et al. 2005; Mescher et al. 2003). Additionally, as explored above, besides changes in social form proportions, other factors may impact gene flow between social forms.

The results showed here are also compatible with the supergene in *S. invicta* working as a selfish genetic element. The social chromosome of *S. invicta* is a putative green-beard gene (Keller & Ross 1998), that is, a genetic element that produces a phenotype that the carriers of such element can recognise. In the context of social behaviour and/or mating, individuals carrying the green-beard genetic element will interact preferentially with each other, leading to the increase in frequency of such element (Hamilton 1964a; Hamilton 1964b; Dawkins, 1978). The Sb variant of the supergene would thus produce some signal that Sb carrying workers could identify, leading to the culling of Sb carrying queens and, potentially, Sb carrying males too (Keller & Ross 1998; Buechel et al. 2014). Some authors have hypothesised the Sb variant did not emerge as a result of selection for the linkage of multiple-queen alleles. Instead, the linkage would have occured between loci involved in the green-beard system, for instance, a locus producing a signal and another one receiving it. The newly formed supergene would work as a drive system, biasing its own spread, through the active elimination of individuals not carrying it. The linkage of the supergene to the multiple-queen colony phenotype would have occurred as a by-product of the supergene formation, or at a later stage (Huang & Wang 2014). The results of the simulations show that the culling of sexuals not carrying the Sb allele is detrimental for multiple-queen colonies, making single-queen colony beneficial alleles are more likely to spread. Therefore, there seems to be a conflict between the spread of the Sb variant and the spread of multiple-queen colony alleles.

Interactions between selfish drive elements and selection for other phenotypes have been reported in other systems. For instance, in the stalk eyed flies there is a drive system in the X chromosome that degenerates the Y chromosome in the eggs, if not counteracted, this results in a bias towards females and sperm limitation (Presgraves et al. 1997). One of the hypotheses explaining the selection for long eye stalks in males is that they could be signaling carriers of resistance to the drive element (Wilkinson et al. 1998). Selection against the drive system could thus results in runaway sexual selection (Rogers et al. 2008).

Limitations

Any model is a simplistic representation of the real world, and the one presented here is not an exception. Some known processes of the life history of *S. invicta* were not modelled here. For instance, queens mating with Sb males have been shown to seek another mate (usually an SB male) (Lawson et al. 2012), the model only accounted for one mating per queen. In the literature, SbSb queens are assumed to be a lethal recessive (Goodisman et al. 2000). In the models without linkage, the proportion of SbSb queens not being able to reproduce was not taken into account. In the model with linkage to the supergene, however, SbSb queens did not reproduce. In natural populations of *S. invicta*, SbSb queens are found in mating flights and in a few nests, even though it is still not clear whether they are viable in the long term (Fritz et al. 2006).

It is also important to note that the estimations of the parameters used here are often extracted from populations in the invasive North American range of *S. invicta*. These populations have a very low genetic diversity due to a founder effect (Ross & Shoemaker 2008), which results in life history changes such as the increase in diploid males produced by multiple-queen colonies (Ross et al. 1993). The invasive population of North America comes from a very specific sub-population from the South American native range (Caldera et al. 2008). Given the high diversity, and in many cases, geographic isolation, of the native *S. invicta* populations (Ahrens et al. 2005; Ross et al. 2007) it is possible that some life history traits relevant for this model may have been poorly estimated.

These issues, however, are unlikely to change the conclusions of this study. They do not affect the qualitative finding that gene flow changes have a strong impact on antagonistic allele frequencies. The discussion centered around the interplay between gene flow changes and allele frequencies would, therefore, remain the same.

The model built for this study uses an analytical, top-down approach. That is, it generalises global behaviours of the system, without taking into account its individual elements. For

instance, the individual selection unit in the model was the colony as a whole. This poses limitations as within-colony interactions are not modelled. These interactions include the potential for inter-caste conflict, for example, where the fitness optima of queens and males would be different (Pennell et al. 2018). Additionally, there are colony-level differences between single and multiple-queen colonies, such as the number, density and fecundity of queens (Vargo & Fletcher 1989). Single-queen colonies, by definition, only have one queen per colony, whereas multiple-queen colonies have tens of egg-laying queens. Consequently, multiple-queen colonies tend to be larger and to produce more individuals than single-queen colonies. This and other demographic and geographic factors were not accounted for in the model. To include them, a bottom-up approach for modelling, such as an individual based model, would have provided complementary insights about any complex emergent properties from simple interactions of individuals. An individual based model would therefore be a good approach to include factors such as caste conflict and geographic and demographic dynamics. Such a model, however, is beyond the scope of this study. The current model could be extended to similar supergene systems, such as that of the white-throated sparrow (Huynh et al. 2011), whereas an individual based model on *S. invicta* would be highly specific to this system, and therefore not applicable to other supergene systems.

Conclusion

Here, we have generated a model to simulate the spread of antagonistic alleles in two social forms of the fire ant *Solenopsis invicta*. These simulations were carried out under four different scenarios: high and low gene flow between social forms with and without linkage of the antagonistic locus to a supergene. The results show that, without linkage, gene flow between social forms has a strong impact on the allele frequencies at equilibrium, where often one of the alleles is lost in the population. The impact of gene flow is far lower when the antagonistic locus is linked to the supergene, with gene frequencies at equilibrium being more stable and often including both the single and multiple-queen beneficial allele. We hypothesis that the supergene acts as a buffer to counteract the impact of gene flow, in an evolutionary strategy similar to bet-hedging. Other mechanisms such as evolutionary conflict or selfish spread of the Sb allele underlying the evolution of the supergene in *S. invicta* are also compatible with our results.

References

Ahrens, M.E., Ross, K.G. & Shoemaker, D.D., 2005. Phylogeographic structure of the fire ant *Solenopsis invicta* in its native South American range: roles of natural barriers and habitat connectivity. *Evolution; international journal of organic evolution*, 59(8), p.1733–1743.

Avril, A. et al., 2019. Asymmetric assortative mating and queen polyandry are linked to a supergene controlling ant social organization. *Molecular ecology*, 28(6), p.1428–1438.

Barney, B.T. et al., 2017. Highly localized divergence within supergenes in Atlantic cod (*Gadus morhua*) within the Gulf of Maine. *BMC genomics*, 18(1), p.271.

Barrett, S.C.H. & Hough, J., 2013. Sexual dimorphism in flowering plants. *Journal of experimental botany*, 64(1), p.67–82.

Bonduriansky, R. & Chenoweth, S.F., 2009. Intralocus sexual conflict. *Trends in ecology & evolution*, 24(5), p.280–288.

Branco, S. et al., 2018. Multiple convergent supergene evolution events in mating-type chromosomes. *Nature Communications*, 9(1).

Buechel, S.D., Wurm, Y. & Keller, L., 2014. Social chromosome variants differentially affect queen determination and the survival of workers in the fire ant *Solenopsis invicta*. *Molecular ecology*, 23(20), p.5117–5127.

Caldera, E.J. et al., 2008. Putative native source of the invasive fire ant *Solenopsis invicta* in the USA. *Biological invasions*, 10(8), p.1457–1479.

Chapman, T. et al., 1995. Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. *Nature*, 373(6511), p.241–244.

Chapman, T. et al., 2003. Sexual conflict. *Trends in ecology & evolution*, 18(1), p.41–47.

Charlesworth, D., 2016. The status of supergenes in the 21st century: recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. *Evolutionary applications*, 9(1), p.74–90.

Charlesworth, D. & Charlesworth, B., 1979. Selection on recombination in a multi-locus system. *Genetics*, 91(3), p.575–580.

Charlesworth, D. & Charlesworth, B., 1975. Theoretical genetics of Batesian mimicry II. Evolution of supergenes. *Journal of theoretical biology*, 55(2), p.305–324.

Clucas, G.V. et al., 2019. Novel signals of adaptive genetic variation in northwestern Atlantic cod revealed by whole genome sequencing. *Evolutionary applications*.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, New York

DeHeer, C.J., 2002. A comparison of the colony-founding potential of queens from single- and multiple-queen colonies of the fire ant *Solenopsis invicta*. *Animal behaviour*, 64(4), p.655–661.

Desjardins, J.K. & Fernald, R.D., 2009. Fish sex: why so diverse? *Current opinion in neurobiology*, 19(6), p.648–653.

Dufresnes, C. et al., 2015. Sex-chromosome homomorphy in palearctic tree frogs results from both turnovers and X–Y Recombination. *Molecular biology and evolution*, 32(9), p.2328–2337.

Dunn, P.O., Whittingham, L.A. & Pitcher, T.E., 2001. Mating systems, sperm competition, and the evolution of sexual dimorphism in birds. *Evolution; international journal of organic evolution*, 55(1), p.161–175.

Fritz, G.N., Vander Meer, R.K. & Preston, C.A., 2006. Selective male mortality in the red imported fire ant, Solenopsis invicta. *Genetics*, 173(1), p.207–213.

Glancey, B.M. et al., 1987. The increasing incidence of the polygynous form of the red imported fire ant, solenopsis invicta (Hymenoptera: Formicidae), in Florida. *The Florida Entomologist*, 70(3), p.400.

Goodisman, M.A., Ross, K.G. & Asmussen, M.A., 2000. A formal assessment of gene flow and selection in the fire ant *Solenopsis invicta*. *Evolution; international journal of organic evolution*, 54(2), p.606–616.

Hamilton, W.D., 1964a. The genetical evolution of social behaviour. I. *Journal of theoretical biology*, 7(1), p.1–16.

Hamilton, W.D., 1964b. The genetical evolution of social behaviour. II. *Journal of theoretical biology*, 7(1), p.17–52.

Hartl, D.L., Clark, A.G. & Clark, A.G., 2007. *Principles of population genetics*, Sinauer

associates Sunderland, MA.

Hecht, B.C. et al., 2013. Genome-wide association reveals genetic basis for the propensity to migrate in wild populations of rainbow and steelhead trout. *Molecular ecology*, 22(11), p.3061–3076.

Hemmer-Hansen, J. et al., 2013. A genomic island linked to ecotype divergence in Atlantic cod. *Molecular ecology*, 22(10), p.2653–2667.

Huang, Y.-C. & Wang, J., 2014. Did the fire ant supergene evolve selfishly or socially? *BioEssays: news and reviews in molecular, cellular and developmental biology*, 36(2), p.200–208.

Hurst, L.D., 1992. Intragenomic conflict as an evolutionary force. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 248(1322), p.135–140.

Huynh, L.Y., Maney, D.L. & Thomas, J.W., 2011. Chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (*Zonotrichia albicollis*). *Heredity*, 106(4), p.537–546.

Innocenti, P. & Morrow, E.H., 2010. The sexually antagonistic genes of *Drosophila melanogaster*. *PLoS biology*, 8(3): e1000335.

Keller, L. & Ross, K.G., 1998. Selfish genes: a green beard in the red fire ant. *Nature*, 394, p.573.

Kirkpatrick, M. & Barton, N., 2006. Chromosome inversions, local adaptation and speciation. *Genetics*, 173(1), p.419–434.

Lawson, L.P., Vander Meer, R.K. & Shoemaker, D., 2012. Male reproductive fitness and queen polyandry are linked to variation in the supergene Gp-9 in the fire ant *Solenopsis invicta*. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741), p.3217–3222.

Lipinska, A. et al., 2015. Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga *ectocarpus*. *Molecular biology and evolution*, 32(6), p.1581–1597.

Mank, J.E., 2017. The transcriptional architecture of phenotypic dimorphism. *Nature ecology & evolution*, 1(1), p.6.

Mescher, M.C. et al., 2003. Distribution of the two social forms of the fire ant *Solenopsis invicta* (Hymenoptera: Formicidae) in the Native South American Range. *Annals of the Entomological Society of America*, 96(6), p.810–817.

Nozaki, H., 1996. Morphology and evolution of sexual reproduction in the *Volvocaceae* (Chlorophyta). *Journal of plant research*, 109(3), p.353–361.

Passera, L. et al., 2001. Queen control of sex ratio in fire ants. *Science*, 293(5533), p.1308–1310.

Pennell, T.M. et al., 2018. Building a new research framework for social evolution: intralocus caste antagonism. *Biological reviews of the Cambridge Philosophical Society*, 93(2), p.1251–1268.

Porter, S.D., 1993. Stability of polygyne and monogyne fire ant populations (Hymenoptera: Formicidae: *Solenopsis invicta*) in the United States. *Journal of economic entomology*, 86(5), p.1344–1347.

Presgraves, D.C., Severance, E. & Wilkinson, G.S., 1997. Sex chromosome meiotic drive in stalk-eyed flies. *Genetics*, 147(3), p.1169–1180.

R Core Team, 2019. R: A Language and Environment for Statistical Computing

Rogers, D.W. et al., 2008. Male sexual ornament size is positively associated with reproductive morphology and enhanced fertility in the stalk-eyed fly *Teleopsis dalmanni*. *BMC evolutionary biology*, 8, p.236.

Ross, K.G. et al., 1993. Effect of a founder event on variation in the genetic sex-determining system of the fire ant *Solenopsis invicta*. *Genetics*, 135(3), p.843–854.

Ross, K.G. et al., 2007. Genetic variation and structure in native populations of the fire ant *Solenopsis invicta*: evolutionary and demographic implications. *Biological journal of the Linnean Society. Linnean Society of London*, 92(3), p.541–560.

Ross, K.G. & Keller, L., 1995. Ecology and evolution of social organization: insights from fire ants and other highly eusocial insects. *Annual review of ecology and systematics*, 26(1), p.631–656.

Ross, K.G. & Shoemaker, D.D., 1993. An unusual pattern of gene flow between the two social forms of the fire ant *Solenopsis invicta*. *Evolution; international journal of organic*

evolution, 47(5), p.1595–1605.

Ross, K.G. & Shoemaker, D.D., 2008. Estimation of the number of founders of an invasive pest insect population: the fire ant *Solenopsis invicta* in the USA. *Proceedings of the Royal Society B: Biological Sciences*, 275(1648), p.2231–2240.

Saddoris, K., Fritz, A.H. & Fritz, G.N., 2016. Evidence of selective mating and triploidy among two social forms of *Solenopsis invicta* (Hymenoptera: Formicidae). *The Florida entomologist*, 99(3), p.566–568.

Sakai, A.K. & Weller, S.G., 1999. Gender and sexual dimorphism in flowering plants: a review of terminology, biogeographic patterns, ecological correlates, and phylogenetic approaches. *Gender and Sexual Dimorphism in Flowering Plants*, p.1–31.

Shine, R., 1989. Ecological causes for the evolution of sexual dimorphism: a review of the evidence. *The Quarterly review of biology*, 64(4), p.419–461.

Sievert, C., 2018. plotly for R. *Available under: https://plotly-book. cpsievert. me*.

Simpson, S.J., McCaffery, A.R. & Haegele, B.F., 1999. A behavioural analysis of phase change in the desert locust. *Biological reviews of the Cambridge Philosophical Society*.

Thompson, M.J. & Jiggins, C.D., 2014. Supergenes and their role in evolution. *Heredity*, 113(1), p.1–8.

Tschinkel, W.R., 2006. *The Fire Ants*, Harvard University Press.

Turner, J.R.G., 1967. On supergenes. I. The evolution of supergenes. *The American naturalist*, 101(919), p.195–221.

Vargo, E.L. & Fletcher, D.J.C., 1989. On the relationship between queen number and fecundity in polygyne colonies of the fire ant *Solenopsis invicta*. *Physiological entomology*, 14(2), p.223–232.

Veltsos, P., Keller, I. & Nichols, R.A., 2008. The inexorable spread of a newly arisen neo-Y chromosome. *PLoS genetics*, 4(5), e1000082.

Widemo, F., 1998. Alternative reproductive strategies in the ruff, *Philomachus pugnax*: a mixed ESS? *Animal behaviour*, 56(2), p.329–336.

Wilkinson, G.S., Presgraves, D.C. & Crymes, L., 1998. Male eye span in stalk-eyed flies

indicates genetic quality by meiotic drive suppression. *Nature*, 391(6664), p.276–279.

Zimmerman, C.E. & Reeves, G.H., 2000. Population structure of sympatric anadromous and nonanadromous *Oncorhynchus mykiss*: evidence from spawning surveys and otolith microchemistry. *Canadian journal of fisheries and aquatic sciences. Journal canadien des sciences halieutiques et aquatiques*, 57(10), p.2152–2162.

Chapter 4: Caste differences in *Solenopsis invicta* are highly tissue specific

Collaborations in this chapter

Dr. Marc Robinson-Réchavi provided the economic resources to produce the data for this chapter

Dr. Ed Vargo provided access to the field samples and helped with the collection and identification of *S. invicta* colonies

Marian Priebe performed the analyses related to specific gene families

I performed the rest of field and lab work and analyses described here and wrote the chapter.

Abstract

Polyphenic traits, where one genotype produces two or more discretely different phenotypes, provide insight into the processes underpinning phenotypic plasticity. More specifically, because polyphenism generates different phenotypes exclusively through gene regulatory processes, understanding its molecular basis can inform our understanding of how gene regulatory networks translate into phenotypes. Here we focus in an extreme case of polyphenism to explore this idea, the morphological castes of the red fire ant *Solenopsis invicta*. This ant species has three different morphological castes: workers, reproductive females and males. All castes share the same genome, the different phenotypes are determined either by feeding during the larval stage (workers and reproductive females) or by ploidy (males and females). We are interested in exploring how gene expression differences are linked to caste differences, and in turn, how this shapes their shared genome. We ask, for instance, how selection affects differently genes that are expressed in different castes compared to genes which are caste-biased. To answer this, we generated tissue-specific RNAseq data from fire ant workers, queens and males. This dataset allows us to study expression patterns in caste-specific tissues, as well as in tissues shared across castes. The results will shed light on our understanding of how polyphenism in particular and phenotypic plasticity in general shape evolutionary trends in the genome.

Introduction

Polyphenism, the phenomenon by which a single genotype produces two or more discretely different phenotypes (Braendle & Flatt 2006), is widely spread across the tree of life. Well known examples include temperature-dependent sex determination in reptiles (Janzen & Paukstis 1991), seasonal forms in butterflies (Simpson et al. 2011) and castes determined by the food allocation to larvae in social insects. Because polyphenism generates different phenotypes by modifying gene regulation, understanding its molecular basis can inform our understanding of how gene regulatory networks produce phenotypes.

Ants provide a paradigmatic example of polyphenism. The vast majority of ant species display different morphological castes, dependent on the sex of the individual. Males form usually a unique reproductive caste, whereas females can develop into reproductive queens or sterile workers.

Hymenopterans have a haplo-diploid system, which means that unfertilized haploid eggs can develop to haploid adult individuals. Typically, haploid eggs will develop into males, and diploid eggs into females (and therefore workers or queens). In several social insects, differences in sexes is determined genetically through a single highly diverse sex determination locus. Whenever an individual is heterozygous for this locus, it will develop as a female (Heimpel & de Boer 2008; Beye et al. 2003). As a result, all adults emerging from unfertilized eggs will be hemizygous for the sex determining locus, and therefore express only one allele, hence becoming males. The differences between queens and workers are environmentally and irreversibly determined during larval development in most species (Miura 2005; Wheeler 1991), for exceptions see (Anderson et al. 2008).

The molecular basis of caste differences can, at least partially, answer more general questions about polyphenism. Indeed, the study of the molecular mechanisms responsible for caste differences has attracted a fair share of research interest (Gadau et al. 2012).

The empirical evidence to date suggests that both highly conserved pleiotropic genes as well as new taxonomically restricted genes with low connectivity seem to be governing phenotypic differences between castes in several social insect species (Mikheyev & Linksvayer 2015; Berens et al. 2015). So far, no particular gene has been found to be a consistent key regulator of caste definition. Instead, whole gene regulatory networks involved in caste differences, seem to be similar across ant species (Morandin et al. 2016). Additionally, theory predicts that gene duplication could play a central role in caste differences, by allowing the subfunctionalisation of gene duplicates in different castes

(Gadagkar 1997; Chau & Goodisman 2017). This hypothesis has received strong empirical support, especially from specific gene families that have undergone several expansions in several ant species (Gadau et al. 2012). For instance, gene families such as odorant binding proteins (OBPs), vitellogenins (Vgs) or chemosensory proteins (CSPs) have expanded in several ant species, and are often involved in differences between castes (e.g. for OBPS: Zhang et al. 2016; Pracana et al. 2017; for Vgs: Morandin et al. 2014; Oxley et al. 2014; Wurm et al. 2011; Corona et al. 2013 for CSPs and other proteins involved in chemical communication: Kulmuni et al. 2013; Koch et al. 2013; Hojo et al. 2015).

Despite the interest in elucidating the molecular basis of caste differences, no clear consensus has emerged on the general mechanisms shaping its evolution. There are several challenges. Namely, the sheer diversity of morphological castes found in hymenopterans (Fjerdingstad & Crozier 2006), and the fact that gene expression between castes seems to be highly life-stage (Ometto et al. 2011) and tissue specific (Abouheif & Wray 2002). Tissue specificity in caste differences is especially relevant, not only because some tissues will play very different roles in different castes (e.g. poison glands in queens and workers -(Vargo 1997; Jackson & Morgan 1993)-), but also because differences in allometry between castes can lead to artefacts in gene expression results if whole bodies are used (Johnson et al. 2013). Most studies to date have focused on either whole bodies or a single tissue from different castes. These shortcomings could have led to inaccurate or incomplete characterisation of gene expression differences.

In this chapter, we generated the most extensive dataset to date in hymenopterans of tissue-specific differences in expression between castes. We extracted RNA from several tissues of adults from three different castes (15 tissues in males, 18 in virgin queens and 17 in workers) of the red fire ant *Solenopsis invicta* (Fig 6.1). Such dataset provides unprecedented resolution to answer questions regarding the molecular basis of caste differences.

Here we show that expression differences between castes are highly tissue-specific; a finding which brings into question the conclusions of previous studies obtained from whole bodies. Caste differences seem to involve thousands of genes, most of which seem to be playing different roles in different castes depending on the tissue where they are expressed. We do find, however, a few genes that seem to be consistently highly expressed in any one caste, albeit with different intensity in each tissue. Analyses focusing on specific gene families of interest including OPBs, CSPs and vitellogenins suggest that many gene duplications exclusive to ants may have been co-opted to play a role in caste differences.

In all, these results provide a valuable case study to understand the regulatory mechanisms underlying the differences between castes in particular, and between discrete phenotypes arising from a single genome in general.

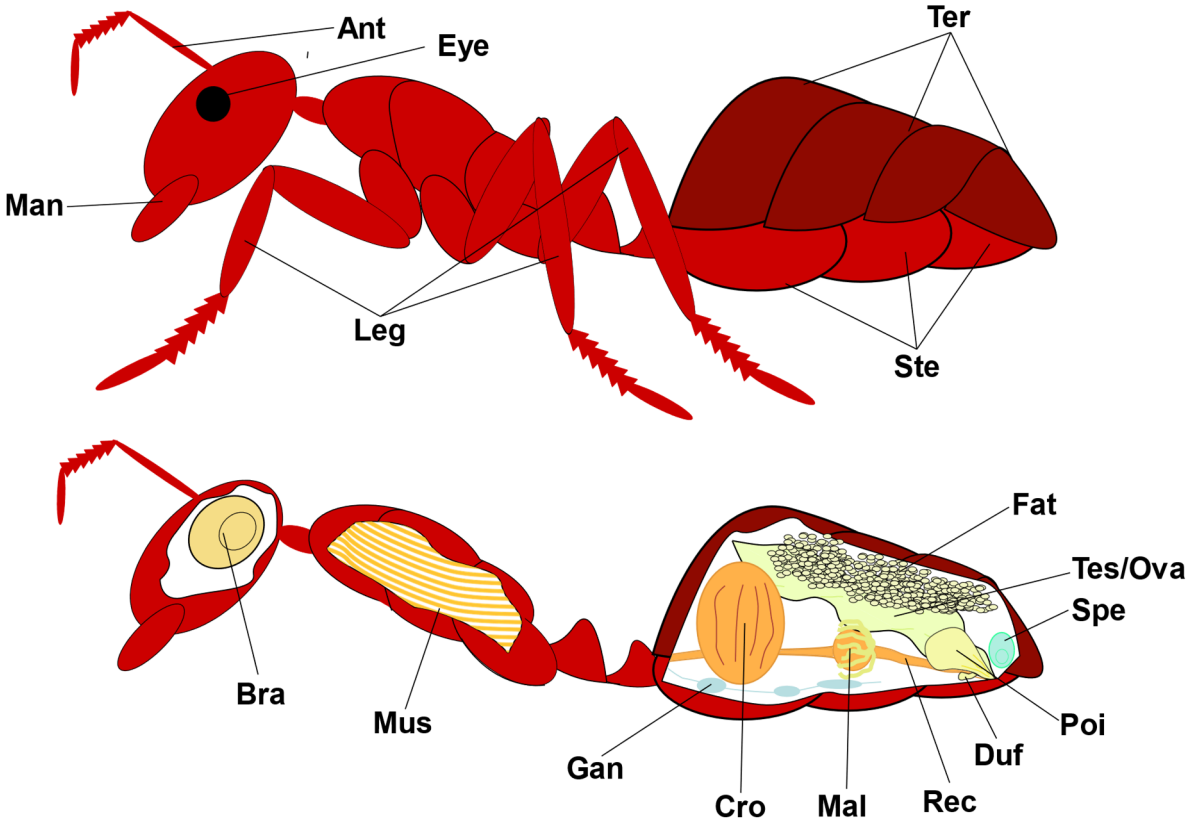


Figure 6.1. Simplified diagram showing the tissues dissected for gene expression from three castes of the fire ant *Solenopsis invicta*. The three code letters correspond to: **Ant**, antennae; **Fat**, fat bodies; **Eye**, eye; **Man,** mandibles; **Leg**, legs; **Ter**, tergites; **Ste**, sternites; **Bra**, brain; **Mus**, muscle (thorax only); **Gan**, abdominal ganglia; **Cro**, crop; **Mal**, Malphigian tubules and midgut; **Rec**, rectum; **Duf**, Dufour gland (queens and workers only); **Poi**, Poison sac and gland (queens and workers only); **Spe**, spermatheca (queens only); **Tes**, testicles (males only); **Ova**, ovaries (queens only). Additionally, we also extracted RNA from the corpse, that is, the rest of the body after dissection.

Materials and methods

Generation of RNAseq data

This experimental design is based on the understanding that the colony rather than the individual should be considered as the unit of biological replication in a social insect. Each replicate was therefore extracted from a different colony, in total the samples were extracted from 15 different colonies. We generated an RNAseq database for several tissues of adult individuals of the red fire ant *Solenopsis invicta* from three different castes: workers, virgin

female reproductives and males. We generated 5 replicates per caste, each replicate comprising a pool of one tissue coming from three individuals of the same colony (except for one female reproductive sample, where only 2 individuals were available).

The samples were collected from colonies in the field near College Station in Texas, USA (GPS coordinates in Annex II, Table AII.1) on the 21st October 2016. After a mound had been identified in the field, the entire colony was transported to the lab and transferred to a plastic box coated in fluon. Once in the lab, the colonies were kept in constant conditions of light and temperature (25ºC) and fed regularly a diet consisting of crickets, apple and sugar water. After at least 8 days of constant conditions, individuals of all available castes from each colony (Annex II, Table AII.1) were collected with soft tweezers and snap-frozen in liquid nitrogen. The samples were then stored at -80ºC. Additionally, workers of each colony were collected and kept in ethanol for genotyping the colony. The whole sampling process was performed on the same day and lasted from 10:30 to 15:30.

The red fire ant *Solenopsis invicta* displays two types of social organisation. This social polymorphism is controlled by a single genetic element (Ross & Keller 1995). For this study we focused only in single-queen colonies. To ensure that RNA was extracted from single-queen colonies only, we extracted the DNA of the workers kept in ethanol using a standard phenol-chloroform method (more details in Annex II, Text AII.1). We then performed the (Krieger & Ross 2002) individual RFLP assays on 9 workers per colony to identify its phenotype. All workers for single-queen colonies are homozygote for this assay, whereas, between 80% and 60% of the workers should be heterozygote in multiple-queen colonies (Buechel et al. 2014).

Once the colonies were genotyped, and their single-queen status confirmed, each individual was dissected in RNAlater straight from the -80ºC storage, to ensure that RNA integrity was preserved at all times. Tissues were extracted into 2mL screw cap tubes containing 1g of ceramic beads over dry ice. Each tube contained a pool of the particular tissue from the three individuals of one caste for a particular colony. 200uL of TRI reagent was added to each tube after dissection and the sample stored at -80ºC before extraction. We performed the dissections following a permuted block design (More information in Annex II, Table AII.1) as to avoid potential confounding factors arising from batch effects. We extracted total RNA using a standard Trizol protocol (More details in Annex II, Text AII.2). The preparation of the 245 individual libraries and sequencing were performed at the Wellcome Genomics Centre in Oxford, UK. Libraries were prepared using the NEBNext Ultra II mRNA kit with an input of 10ng or maximum available. 16 PCR cycles were used cycles for indexing and amplification. An equimolar pool of samples was sequenced in two lanes of an Illumina HiSeq4000

sequencer at 75bp paired-end and 16 lanes (2 flowcells of 8 lanes each) an Illumina HiSeq4000 sequencer at 50bp single end reads. The sequencing produced an average of 1,382,570 reads per sample, with a maximum of 2,921,715 reads and a minimum of 699,318 reads. A replicate for female reproductive corpse tissue and a replicate for male fat tissue did not produce material enough and were only sequenced in the two lanes of 75bp paired-end.

RNAseq quality check

Raw reads and mapping quality checks

We removed the Universal Ilumina and Tru Seq adapters from the raw RNAseq reads using Cutadapt (v1.13; Martin 2011)) with default parameters. Because major sequencing biases typically affect all the samples in an entire lane, or an entire library across all lanes, we performed quality control at both levels. For this quality control, we merged the raw reads by sequencing lane or by sample of origin quality checked them using FastQC (v0.11.5; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We then mapped the reads from each sample and lane individually to the *Solenopsis invicta* genome reference (gnG assembly, release 41 in Ensembl Metazoa) using STAR (v2.5.3a; Dobin et al. 2013). We first generated an index using the 'sjdbGTFtagExonParentTranscript=Parent' option. We then mapped the reads twice using the 'out.tab' file for the second run, and set 'sjdbOverhang' option to 49, as recommended by the developers. We merged the resulting alignment files (BAM format) by sequencing lane or by sample using Samtools merge (v 1.9; Li et al. 2009)), and then quality checked them using Qualimap rnaseq (v2.2.1; Okonechnikov et al. 2016). The results of these quality checks were integrated together using MultiQC (v1.5; Ewels et al. 2016).

Overall, these quality checks show no major effect by lane. There are, however, some patterns which emerged when the reads were grouped by sample. For instance, some thoracic muscle samples, particularly those belonging to males and female reproductives, have a very high percentage of repeated sequences (in one case, going up to 55% of the total reads). These were mitochondrial sequences. This pattern might be expected in these samples, since muscle is enriched in mitochondrial expression, especially in winged individuals, where thoracic muscles need to cope with the energetic cost of flying. Mitochondrial sequences in other tissues could either belong to actual mitochondrial expression or to unannotated nuclear mitochondrial genes (numts). Interestingly, all brain samples appear to have a higher percentage of reads mapping to intergenic and intronic

regions. This effect is not caused by a few highly expressed genes being poorly annotated, since this pattern holds even when highly expressed genes are not considered (data not shown).

A few of the samples, however, seem to perform poorly in many metrics. Specifically, 8 samples from one worker replicate show a very low percentage of reads aligning to exonic regions, a poor 3'-5' bias coverage and tend to have a relatively low number of mapped reads in relation with the number of total raw reads. This may be indicating that RNA was slightly degraded for this samples. The case for discarding these samples was further evaluated using PCA analysis (see below).

Principal component analyses (PCAs)

In addition to the aforementioned quality checks performed on the raw reads and the alignments, we also performed several principal component analyses (PCAs) on the read counts per gene of the RNAseq data. All the PCA plots were generated using the R package 'ggbiplot' (Vu 2011).

These analyses are expected to produce biologically meaningful patterns (e.g. clustering by tissue or caste), and should expose any potential technical batch that could affect downstream expression analyses. The read counts for the PCA were obtained running the Kallisto quant (v0.44.0, (Bray et al. 2016) tool with standard options for single end reads. The mean insert size was set to 300bp, with a standard deviation of 20bp. We used the *S. invicta* cDNA reference from Ensembl Metazoa (gnG assembly, release 41). We generated a Kallisto output for the reads generated per sample and per lane in the sequencer. These reads were then imported into R using Tximport (v1.2.0; Soneson et al. 2015) and DESeq2 (v1.14.1; Love et al. 2014). The counts were normalised by library size using the DESeq method. We aggregated the raw counts by adding the read counts per gene by sequencing lane or by sample. We plotted the first two PCs of the PCA by sample as they explained most of the variance (Fig AII.1). The combination of PC 1 and 2 (Fig AII.2), shows that eight samples from a worker replicate may be problematic, because they form a separate cluster. These samples are the same that performed poorly for many of the other quality check metrics. These eight samples show, in general, more read counts for highly expressed genes than the same tissues in other worker replicates, but no expression for genes that have low expression in other replicates. This may be indicating that the RNA used for generating these libraries was at least partially degraded. Because there is no reason to think that RNA degradation would be biased against a particular sequence, it will look as if the surviving (mostly highly expressed) sequences look as if they were relatively highly

expressed compared to other replicates. Consequently, these 8 samples were removed from downstream analyses. Other than this clustering attributed to RNA degeneration, there seems to be a more welcome strong biological signal: samples seem to cluster by tissue and caste. For instance, most tissues tend to cluster together. Caste has also an effect in the clustering, albeit fainter, with males and females clustering together, but worker samples being more scattered (Fig AII.3). Indeed, when grouped by RNA extraction batch (Fig AII.4), there is no clear clustering pattern, suggesting that the RNA extraction process did not affect the perceived expression patterns. Despite the quality check of the raw counts and alignment per lane showing no clear effect, the PCA with samples merged by lane (Fig AII.5) shows that normalised read counts per gene do group by flowcell in the sequencer, which implies that sequencing run has an effect on the perceived expression patterns. This effect, is relatively small, as shown by the fact that the variance explained by these PCs is small and similar across PCs (Fig AII.6). Indeed, when a PCA is performed in all raw read files independently (Fig AII.7), the effect of lane or flowcell is very small in comparison with other effects discussed below. Nevertheless, flowcell effect is accounted for as batch effect in further analyses.

Gene expression analyses

Generation of read counts per gene

We estimated the number of significantly differentially expressed genes between castes within tissues. In light of the results of the QC (above), we obtained the read counts merging the read files by flowcell. That is, we generated 4 different read count files per sample: 2 per each run at 75bp paired-end and 2 per each run at 50bp single-end. We then obtained the read counts per gene using Kallisto with the same options as described above for the single-end reads and with the standard options for paired-end reads. We also used the same reference as described above, with the exception that the mitochondrial genes were removed. The read counts were then imported into R using tximport and we obtained the transcript-gene equivalence table from Ensembl using Bioconductor package biomaRt (Durinck et al. 2009). The 8 samples deemed as of low-quality were removed from any downstream analyses.

Gene expression differences for tissues common to all castes

We first analysed the gene expression differences between castes within tissues. For this analysis, we only focused on tissues present in all castes. That is, we did not include in this

analysis the poison and Dufour glands, ovaries, spermatheca and testis; after these restrictions, 14 tissues remained for this analysis. The significance level of gene expression differences was estimated running limma-voom (Ritchie et al. 2015) with standard parameters (Law et al. 2018). This analysis tool is part of the Bioconductor package edgeR (Robinson et al. 2010). Before the analysis, we filtered out genes with a median counts per million (CPM) across all samples of less than 0.5. 4479 genes with were removed from analysis, leaving 10,430 for the downstream analysis. We then produced a nested design (caste within tissue), using flowcell as a blocking factor. We generated a pairwise contrast between caste per tissue (3 contrasts in total per tissue) to obtain gene expression differences. We considered a gene to be differentially expressed in each comparison when It produced a Benjamini-Hochberg corrected p value of < 0.05 and a log2 fold change difference in expression between castes of more than 1 (*i.e.* 2-fold change in expression).

Gene expression differences for caste-specific tissues

The tissues that were not present in all castes were analysed separately. The poison and Dufour glands are present in both queens and workers. They were analysed following the same steps as those described in the previous section, with the exception that the only comparison performed was queens against workers.

The ovaries and the testes are caste-specific reproductive tissues present only in queens and males respectively. We compared directly the gene expression level differences between testes and ovaries using limma-voom with the same parameters as described above.

Because the spermatheca is only present in queens, we used limma-voom only to test which genes were expressed to a level significantly different from 0 in this tissue. In other words, we tested which genes were expressed at all in this tissue.

All the plots for this section were generated using ggplot2 (Wickham 2016) and ggtern (Hamilton & Ferry 2018).

Analysis of specific gene families

We analysed the expression patterns for three specific gene families, namely Vitellogenins (Vgs), odorant binding proteins (OBPs) and chemosensory proteins (CSPs) in *S. invicta*. These gene families were chosen because of their potential relevance in caste differences and because accurate gene models are available for them. Based on the literature, we

updated the gnG reference of *S. invicta* with the improved gene models. The location of the Vg genes in the reference were taken from (Wurm et al. 2011). Vg3 and Vg2 was already present in the original reference. Because Vg1 and Vg4 were merged as a single gene in the original reference, we simply split it manually to obtain all the Vgs. The sequences of the CSPs were obtained from (Kulmuni et al. 2013). We then used the best matches from exonerate (v2.4.0; Slater & Birney 2005) by using the gnG assembly as a reference and the CSP sequences as queries to extract the positions of the CSPs in the *S. invicta* genome. The positions of the OBPs in the gnG reference were obtained from (Pracana et al. 2017). The positions of both the CSPs and OBPs were then used to generate an updated gff annotation for *S. invicta*, by replacing the previous gene models for Vgs, CSPs and OBPs with the new positions, while also retrieving the rest of gene models in the genome. We then generated an updated reference for Kallisto by extracting the sequences of all genes using the gff annotations and the gnG reference assembly for *S. invicta* using genometools (v1.5.9; Gremme et al. 2013). We then ran Kallisto using the parameters described above for paired end reads using the updated gene sequences as a reference and the read files merged by sample. The results were plotted using the R packages ggplot2, and dendextend (v1.12.0; Galili 2015).

GO terms enrichment analyses

A gene ontology (GO) enrichment analysis was performed merging all the counts by individual, irrespective of tissue. This allowed a GO terms enrichment analysis on the gene expression differences between castes across all tissues. The GO term enrichment was performed using the Bioconductor (Huber et al. 2015) package TopGO (v2.36.0; Alexa & Rahnenfuhrer 2016) in R (v3.6.1; R Core Team, 2019). TopGO was run using the algorithm “weight01” and the enrichment tested using the “ks” test, using the p values from the pairwise comparisons as scores. The GO terms per gene were obtained from Ensembl Metazoa using the Bioconductor package biomaRt (v2.40.1; Durinck et al. 2009).

Results

Most genes have a caste-tissue effect

Out of the 10,430 genes analysed, we found 135 genes which showed at least one caste difference between castes in all tissues, of these, 24 were always differentially expressed between queens and workers, 41 between males and queens and 42 between males and workers. Assuming that these patterns are a proxy for overall differences between castes, they show that differences within reproductive status (males vs queens) are bigger than within sex (workers vs queens). The results of the general GO terms enrichment analysis between castes across tissues (available in Online Annex II at http://bit.ly/OnlineAII) show several terms that appear in all comparisons, including “translation” (GO:0006412), “oxidation-reduction process” (GO:0055114), or “nucleoside metabolic process” (GO:0009116). Additionally, processes involved in lipid metabolism such as “lipid transport” (GO:0006869) or “lipid metabolic process” (GO:0006629) were present in at least two of the comparisons. These biological processes and related cellular component categories such as “ribosome” (GO:0005840) are thus likely to play an important role in caste differences.

Tissues have varying levels of caste-specific differences

The total number of differentially expressed genes between castes in different tissues is variable (Fig 6.2, see also Fig AII.8). The most marked differences were found in antennae and fat bodies. Antennal tissue had many more differentially expressed genes between males and females (workers or queens; 1881 and 1601 respectively) than between workers and queens (453). Hence, by this criterion, it was the tissue showing the most differentiation between the sexes, irrespective of reproductive status. Such a pattern could be indicating that queens and workers antennae are more functionally diverse than those of males. Fat body showed a different pattern. In this case queens had more differentially expressed genes when compared to the other two castes (3399 for queens vs workers and 3324 for queens vs males) than the comparison between males and workers (1917). These numbers are consistent with the different roles of fat stores in the different castes. For instance, both males and queens need long-term storage for flight, as opposed to workers. Additionally, the fat tissue in insects are involved in the production of proteins and metabolites which are likely to vary across castes such as vitellogenin, which is known to be expressed mostly in queens (Arrese & Soulages 2010). Malphigian tubules and midgut have the lowest differentially expressed genes for any comparison (361 genes differentially expressed

between queens and workers), consistent with similar amounts of digestion/detoxification taking place in different adult castes

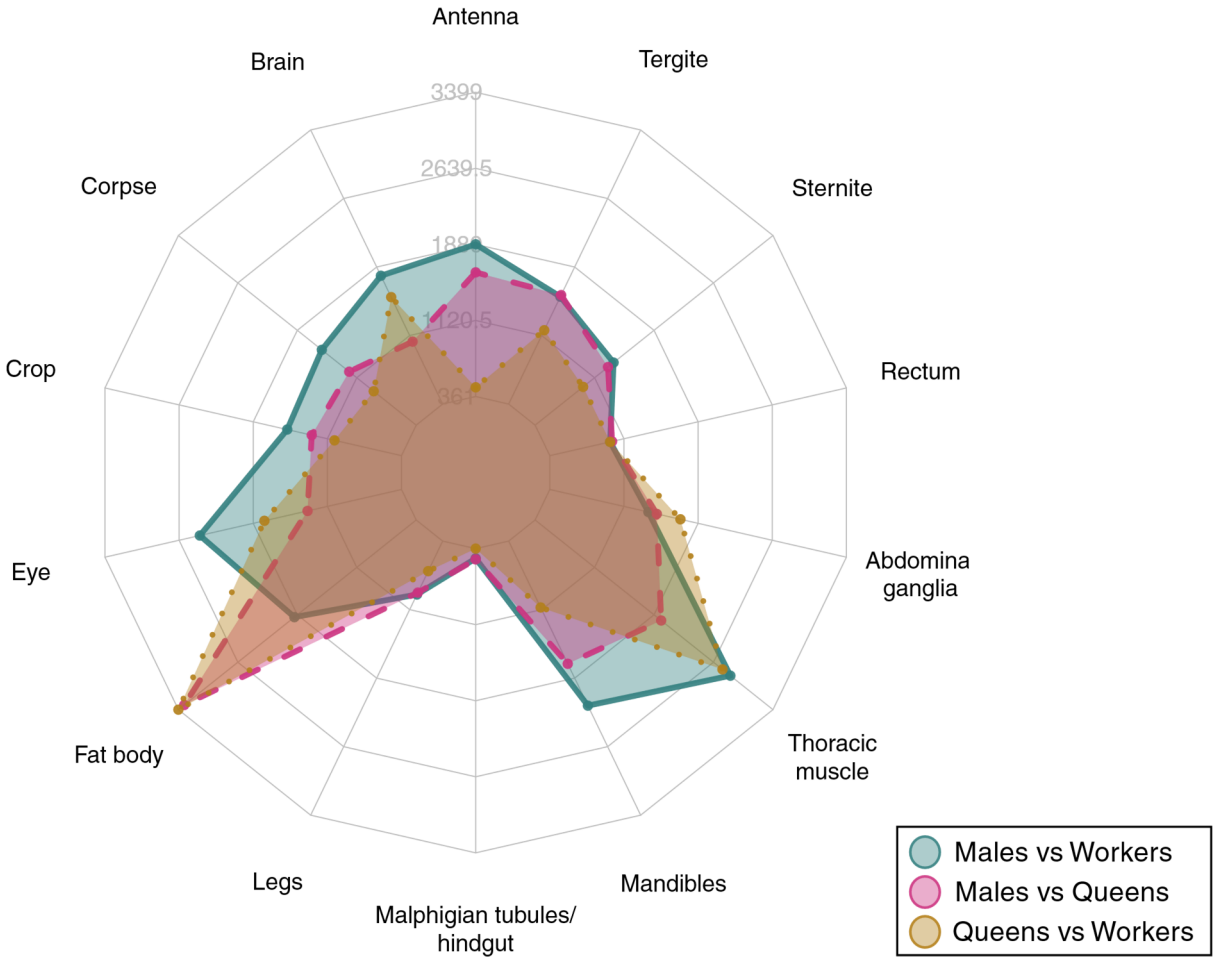


Figure 6.2. Number of differentially expressed genes in all tissues common to all castes for each comparison. The total number of differentially expressed genes range from 361 (minimum number of differences in the Malphigian tubules/midgut) up to 3399 (maximum differences in Fat body). The area of the polygon for each comparison is a measure of the total number of differences across tissues.

Fewer tissues have genes with male-biased expression, but fewer genes have queen-biased expression

Differences among tissues

The analysis of the biases in expression, broadly showed equity among the castes – approximately one third of the genes showing significant differences in expression-level were most highly expressed in each caste. There were, however, small but discernible differences among tissues (the points in Fig 6.3, are displaced from the centre of the triangle towards an apex corresponding to a particular caste). The vast majority of tissues had more queen-biased genes (Antenna, Malphigian tubules/midgut, Thoracic muscle, Rectum, Legs, Brain, Mandibles, Eye, Corpse), the rest had a slight bias towards either workers (Abdominal ganglia, Sternites and Tergites) or males (Fat body and Crop). The most biased tissues were Antenna for queens (56% of differentially expressed genes are queen-biased), Crop for males (37% of differentially expressed genes are male-biased) and Sternite for workers (39% of differentially expressed genes are worker-biased).

Differences among loci

When the patterns for each gene are averaged across tissue, however, there were fewer strongly queen-biased genes, and more genes highly expressed in males and workers (shown as points displaced from the centre of Fig AII.9) towards the corresponding apex). The genes with the highest expression level in each caste (relative to the other two) were: “vitellogenin-2” for queens (LOC105205782), “fatty-acid amide hydrolase 2-B-like” for males (LOC105204825), and “pheromone-binding protein Gp-9” (LOC105194487) for workers.

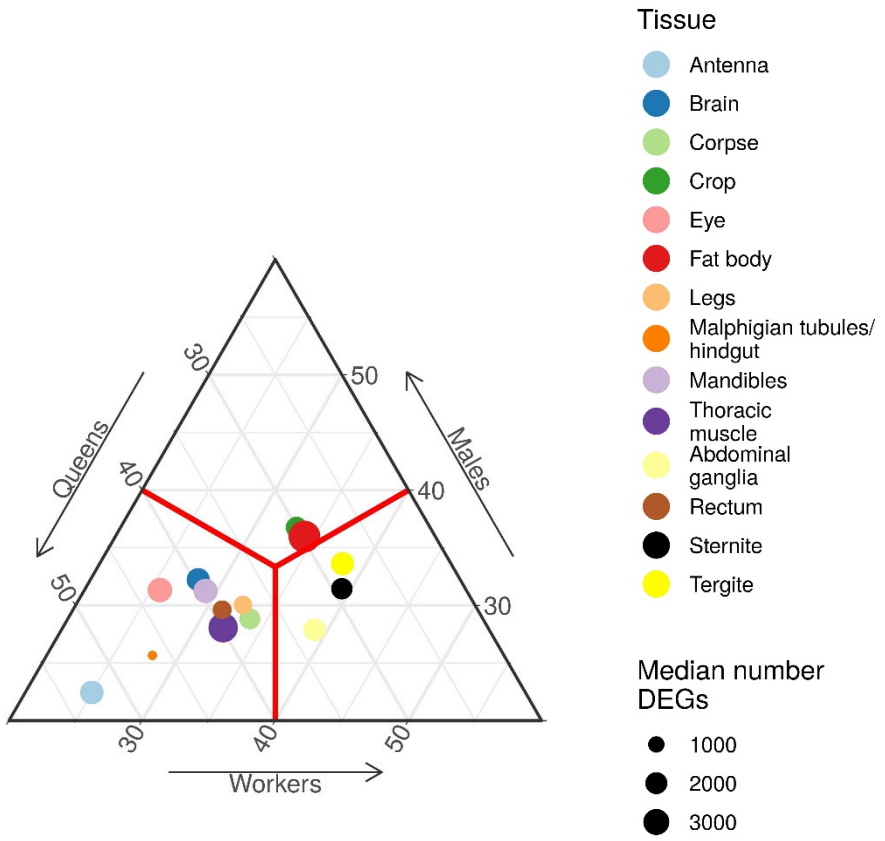


Figure 6.3. Expression bias in each tissue. Each point represents a tissue, and its location shows the bias towards expression in different castes. Most tissues cluster around a central point with an equal proportion (33⅓ %) of genes biased towards each caste ('biased' here indicates genes with significantly higher expression in that caste). Points displaced toward an apex have more genes biased towards the corresponding caste. The points closest to each apex are separated by red lines. The scale along each edge of the triangle shows percentage of biased genes in the comparison between a pair of castes.

Groups of genes within gene families show expression patterns consistent with subfunctionalisation

We analysed the expression patterns across all tissues for 3 gene families: vitellogenins (Vg), odorant binding proteins (OBPs) and chemosensory proteins (CSPs). We found that in all three cases, groups of genes within the family displayed different expression patterns. This pattern is indicative of genes performing different functions within gene families.

Vitellogenins (Vgs)

S. invicta has 4 types of vitellogenin (Vg1, Vg2, Vg3, Vg4) as a result of a series of gene duplication events. We find that Vg2 and Vg3 were highly expressed across all tissues in

queens but not other castes (Fig 6.4a). Expression levels of both of these genes were highest in the fat body and tergite and lowest in the ovaries.
Expression of Vg1 was present in all three castes being highest in the workers and lowest in the males. The fat body, tergites and sternites of workers had the highest tissue-level expression. Expression of Vg4 had high expression levels in all three castes in the antennae, brain and abdominal nerve ganglia; the highest expression was in males.

Chemosensory Proteins (CSPs)

There are 21 CSPs in the fire ant genome. We find some CSPs with antenna biased expression in our dataset (CSP1, CSP7 in workers and queens & CSP19)(Fig 6.4b).

Ten of the CSPs (CSP8, CSP9, CSP10, CSP12, CSP13, CSP15, CSP16, CSP17, CSP20 & CSP21) were expressed at high levels across many of the external, cuticular tissues sampled, particularly in the sternite, tergite and legs but in some cases also in the mandibles (CSP8, CSP9, CSP12, CSP16, CSP17 & CSP21), antenna (CSP9, CSP12, CSP15, CSP17 & CSP21) and the crop (CSP15). This general expression pattern was usually found in all castes but in some cases there was slightly higher expression in workers (CSP8, CSP16, CSP17, CSP20 & CSP21) or in male antenna (CSP17 & CSP21).

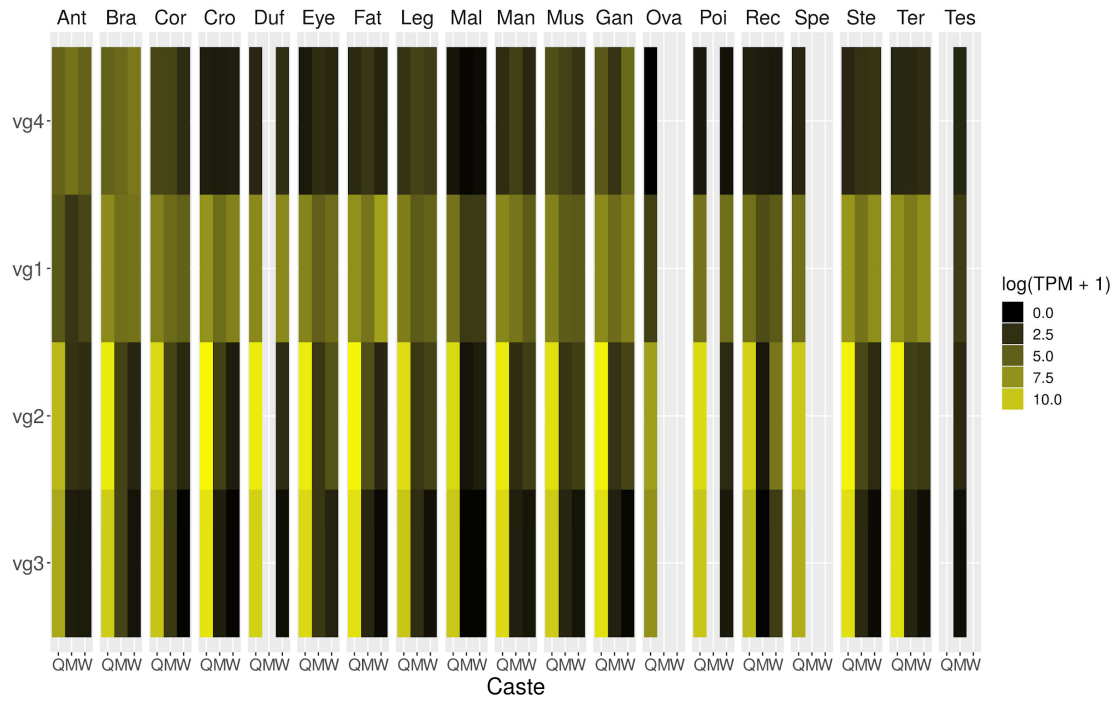
Other CSPs show a wide variety of expression patterns across our dataset including high expression in the rectum of all castes (CSP2), high expression in the spermatheca (CSP3 & CSP5) and high expression in the ovaries (CSP4 & CSP6).

Odorant Binding Proteins (OBPs)

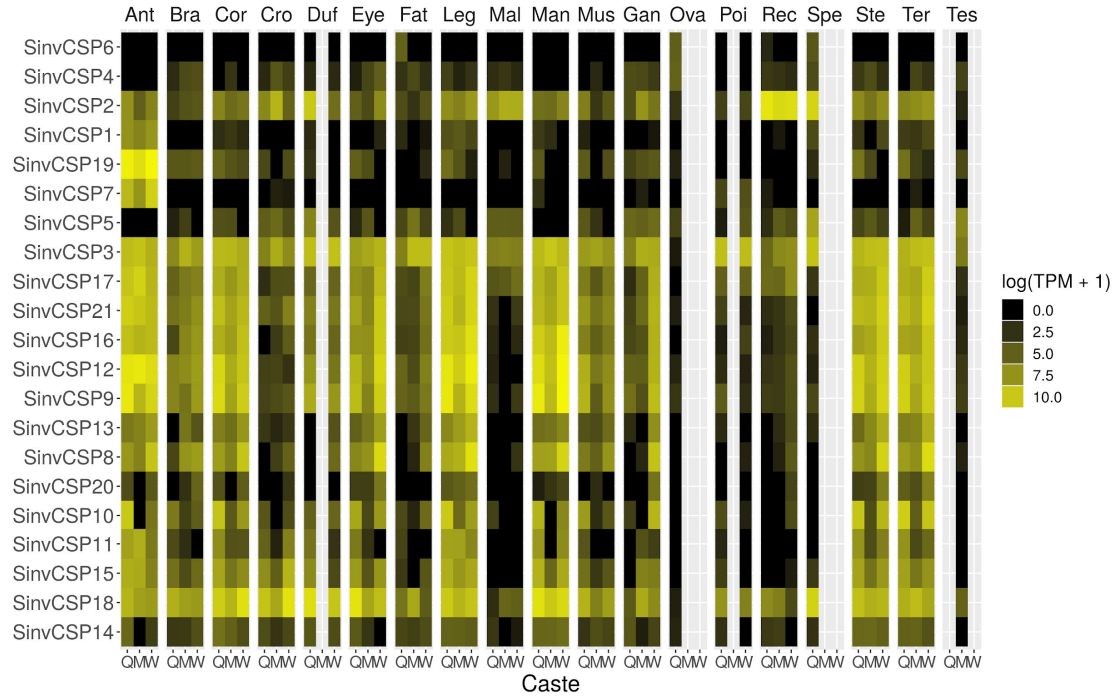
The fire ant genome contains 24 annotated OBPs, 9 of which are within a supergene region subjected to low recombination and linked to different forms of social organisation within this species.
In our analyses we found two common patterns of expression amongst the OBPs (Fig 6.4c). Eight of the OBPs showed tissue specific expression in the antennae (OBP1, OBP2, OBP5, OBP6, OBP11, OBP14 & OBPZ1), brain (OBP10 & OBP14) and abdominal ganglia (OBP10 & OBP14) in all castes.
Seven of the OBPs (OBP3/Gp9, OBP4, OBP12, OBP13, OBP15 & OBP16) showed a higher expression in workers and queens across all tissues, with highest expression in the fat body. These OBPs not only cluster together based on their expression profiles, but in a phylogenetic analysis of their DNA sequence they also form a monophyletic cluster, which, in addition corresponds with their location in the supergene of *S. invicta* (Fig 6.5).

The rest of the OBPs also show expression patterns that are not necessarily consistent with genes involved in olfaction. OBP9 for example is only expressed in the testes of males while OBPZ2 is only expressed in the spermatheca of queens.

a)



b)



c)

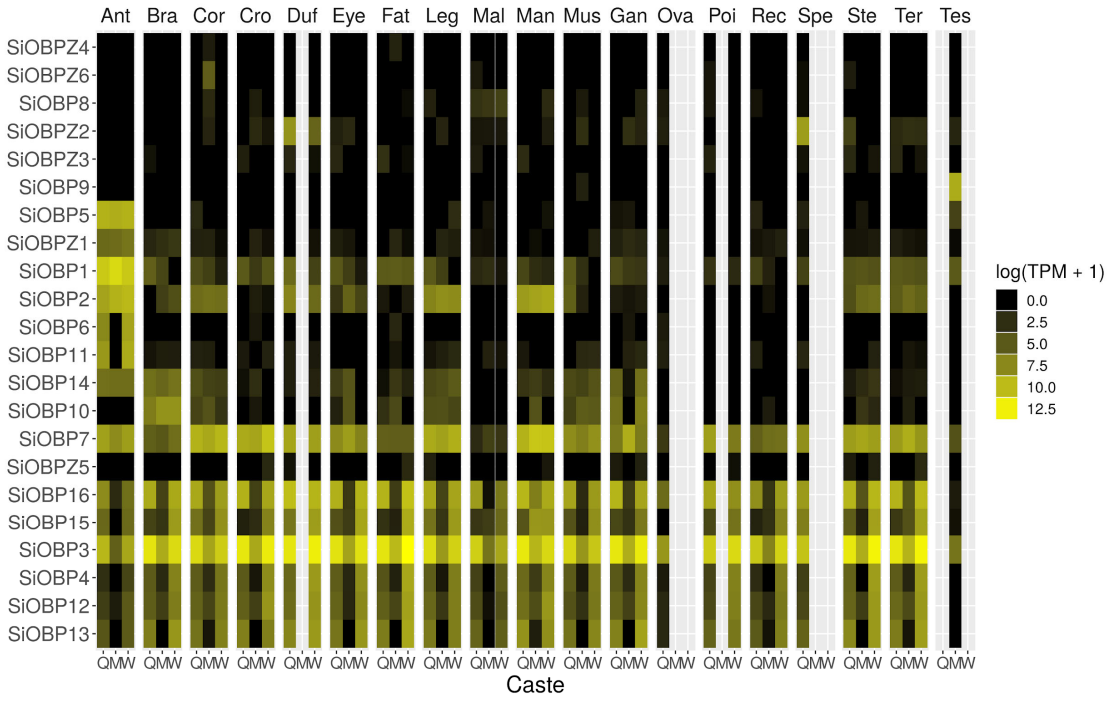


Figure 6.4: Heatmap showing the expression patterns of all **a)** vitellogenins (Vgs), **b)** chemosensory proteins (CSPs) and **c)** odorant binding proteins (OBPs) across body parts (top axis) and castes (bottom axis). Each row represents one gene, each individual cell is a combination of gene, caste and tissue. Gene expression is measured as the logarithm of transcripts per million (TPM) + 1 for each gene. That is, the logarithm of the relative abundance of transcripts from each gene compared to the total of RNA molecules in the dataset. Note that some cells are empty, where the tissue was not available for a specific caste. The genes are ordered by similarity in expression patterns, that is, genes with similar expression patterns overall are plotted next to each other. The three code letters for tissue correspond to: **Ant**, antennae; **Fat**, fat bodies; **Eye**, eye; **Man,** mandibles; **Leg**, legs; **Ter**, tergites; **Ste**, sternites; **Bra**, brain; **Mus**, muscle (thorax only); **Gan**, abdominal ganglia; **Cro**, crop; **Mal**, Malphigian tubules and midgut; **Rec**, rectum; **Duf**, Dufour gland (queens and

Discussion

We generated the most comprehensive tissue-specific RNAseq dataset for any hymenopteran to date. The first analyses resulting from this dataset paint a complex picture of the gene expression differences between castes of *S. invicta*. More specifically, we show that caste differences are highly tissue specific, with some tissues seemingly more specialised in particular castes. Additionally, the results here show that some gene families that have undergone several lineage specific duplications show signs of playing different roles in different castes. In all, these results are a big step forward towards the unravelling of the molecular machinery involved in caste differences.

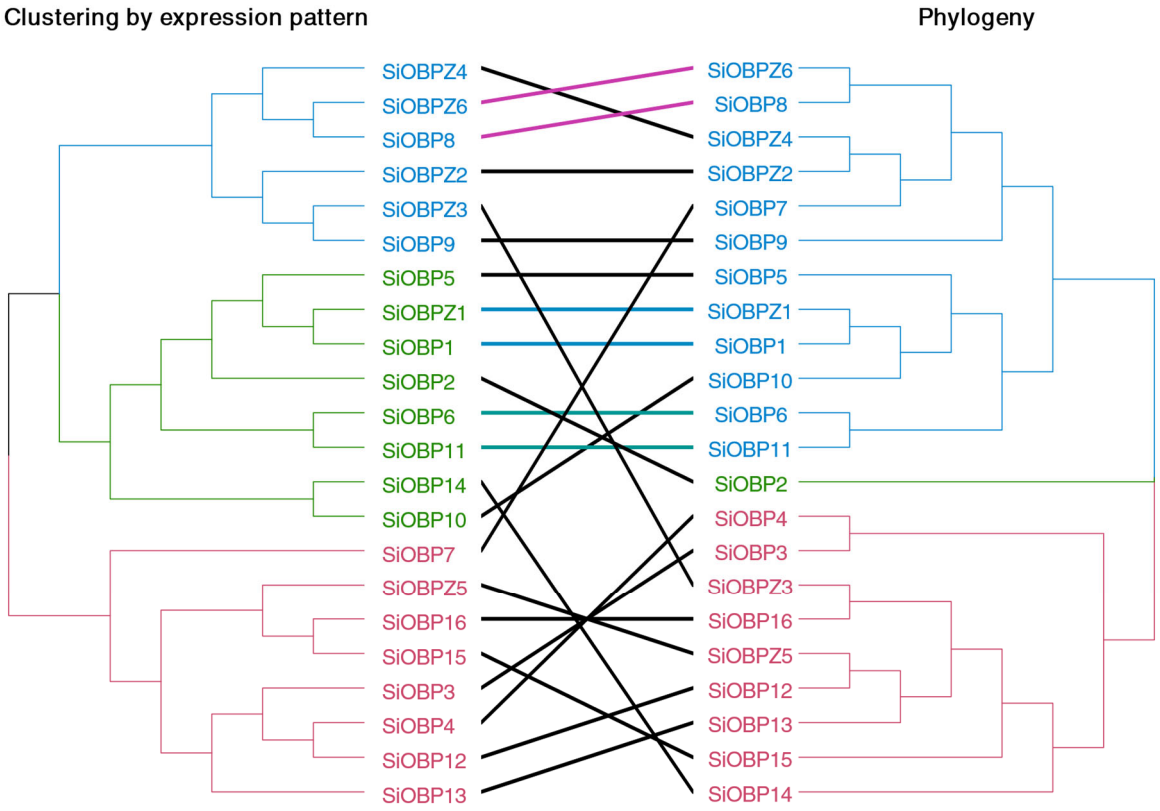


Figure 6.5: Equivalence of clusters of odorant binding proteins (OBPs) based on expression patterns (left) or on phylogeny (left, phylogeny based on the results from (Pracana et al. 2017). The clustering by expression patterns forms three clusters, whereas phylogenetically there are two big clusters of OBPs. Of these two clusters, the bottom one (pink) corresponds to all OBPs linked together in the supergene of *S. invicta*. This phylogenetic cluster corresponds roughly (with the exception of SiOBP14 and SiOBPZ3) with the bottom (pink) cluster by expression patterns. This pattern suggests that phylogenetically close genes have similar expression patterns across different body parts and castes, and, therefore, potentially similar functions.

Across tissues general expression patterns in caste differences

Our results emphasise the importance of considering tissue-specific data when it comes to test for gene expression differences between morphologically diverse individuals. Most tissues had thousands of genes being differentially expressed between castes. Conversely, only a few (135) showed caste differences in all tissues and none showed consistent differences between the three castes in all tissues.

These findings are congruous with previous work performed in honey bees, in which multiple genes were found to be involved in division of labour within workers, rather than a single "caste gene" (Johnson & Jasper 2016). These caste-biased genes are, additionally, likely to be pleiotropic, given that they are expressed in all body parts. This result, again, supports previous findings using tissue-specific data from honey bees which suggest that caste is defined by highly pleiotropic genes (Jasper et al. 2015). The few hundred genes found here consistently differentially expressed between castes could be used as an initial list of candidate genes that could be acting as "master genes" for caste differences. Such a list would not be complete, however, unless different developmental stages were also considered, since expression biases between castes are known to appear at different stages in ants (Ometto et al. 2011; Morandin et al. 2015).

When averaging expression across tissues, we found a large number of genes showing expression differences between males and workers, fewer between males and queens, and least between workers and queens. These patterns are consistent with the fact that workers and queens share sex, males and queens share reproductive status but workers and males share neither. Another pattern is that workers have more highly expressed genes compared to males and queens. This pattern would be a consequence of the higher diversity of environments that workers face compared to the other castes (Feldmeyer et al. 2014). Both these patterns are consistent with previous work using whole-body expression data (Ometto et al. 2011) and are, therefore, what we would expect when considering the expression patterns averaged across all tissues.

The general GO terms analysis, considering all tissues together, reveal that some functional groups seem to be involved in all caste differences. Specifically, "translation" and "translational elongation" (GO:0006412 and GO:0006414), "oxidation-reduction process" (GO:0055114) and processes related to lipid metabolism (GO:0006629 and GO:0006869). The later two terms may be explained by both queens and males having the ability to fly, which requires a high energetic demand (Helms, 2018). Lipid metabolism could also be indicating differences in cuticular molecules, used for chemical communication in social organisation, which are tightly linked to lipid metabolism (examples are reviewed in (Richard

& Hunt 2013). To further this argument, “sensory perception of taste” (GO:0050909) is one of the top 3 terms enriched in the comparison between queens and workers. This term is associated with the neurological processes involved in the processing of environmental chemicals. In other words, the seven genes associated with it are likely to play a role in chemical perception. Finally, the fact that translation is a process enriched in all comparisons could be pointing at the relevance of alternative splicing as a molecular mechanism underlying caste differences. It is important to note that alternative splicing has been shown to play an important role in caste differences in other social insects (Foret et al. 2012; Price et al. 2018; Glastad et al. 2016).

Our tissue-specific data reveals important patterns that are not seen in whole body studies, or in our aggregated data (averaged across tissues). Even though workers and males have a larger proportion of highly expressed genes, the results here show that queens had a single gene (Vitellogenin 2) which was consistently highly expressed across all tissues, and had negligible expression in the other two castes. Neither males nor workers had a comparable gene with higher expression across all tissues. The higher expression of Vitellogenin 2 in *S. invicta* queens was already known (Wurm et al. 2011), but only with the tissue-specific data used here does the exclusive nature of this bias become clear. In other words, our results support the idea of a potential “queen gene” (understood here as a gene consistently linked to the queen phenotype, but not necessarily causal), but not that of “worker” nor “male” genes, and certainly not a “caste gene”.

Tissues show different levels of caste specificity

Further patterns emerge from the analysis of the different tissues. For instance, most tissues in common between all castes have more genes exhibiting a bias towards females (i.e. showing significantly higher expression in queens or workers). This pattern was not seen when expression is averaged over tissues, on the contrary we have seen, in those data, there were fewest genes biased toward queens. These expression patterns suggest that some tissues are more specialised to perform a particular task in a specific caste. For instance, the eye has many queen and male biased genes compared with other tissues and very few worker biased genes. These general expression patterns can be linked to the different functions of the eye in each caste. In workers, the eye is small compared to the other two castes. Males, on the other hand, need accurate vision to be able to identify females in the mating flight (Shik et al. 2013). As a result, they tend to have larger, more complex, eyes – a pattern which can be explained by the importance of vision in males, for which fitness depends largely on successfully finding a mate (Narendra et al. 2011). A comparable argument can be made for queens. Even though mature egg-laying queens

spend most of their lives underground (where vision is not critical for their survival), the queens used here were virgin alates. These are queens that have yet to find a mate in a mating flight, and therefore refined vision would be required to find a mate and establish a colony (Shik et al. 2013). The antenna tissue show a completely different pattern: far fewer genes were highly expressed in males than in either queens or workers. Again this pattern is consistent with what can be inferred from the morphology and histology of the tissue and the associated behaviours: male antennae are shorter in *S. invicta* than those in the other castes. The antennae of both queens and workers perform a wider range of functions, which presumably explains why male antennae have fewer antennal sensilla (Nakanishi et al. 2009) and fewer antennae-specific muscles (Ehmer & Gronenberg 1997) across a number of ant species. In short, male antennae are specialised in fewer tasks than those of other castes. The examples of eye and antennae suggest that the number of highly expressed genes can be used as a measure of the degree of specialisation of a tissue in a specific caste.

Using this measure, the least specialised tissues are the midgut and Malphigian tubules, that is, organs that deal with nutrient reabsorption and detoxification (reviewed in (Phillips et al. 1987; Dow 2009). The most differentiated tissues are the fat bodies. This result is credible, since arthropods' fat bodies synthesise hormones and pheromones that are then released in the haemolymph (reviewed in (Howard & Blomquist 1982). Different castes produce different types of pheromones (Hölldobler et al. 1990). Additionally, males and virgin queens need to store large amounts of energy for flying in the fat bodies (examples in Helms, 2018). Finally, queens need additional energy input for egg production (Tschinkel 1993). All these different functions of the fat body could contribute to the highly differentiated expression patterns described here.

Expression patterns can also be used to obtain additional information on tissues for which the function is incompletely known in *S. invicta*. For instance, the Dufour gland is known to produce trail pheromone in workers (Vander Meer et al. 1988). It may also be involved in producing queen pheromone in queens (Vargo & Hulsey 2000), but to date, its specific function in queens is unknown. The expression results show more highly expressed genes for queens in the Dufour gland than for workers. That would imply that the Dufour gland in queens is more specialised or, at least, is performing more diverse functions than in workers. Dufour glands are involved in the production of sexual pheromones for male attraction in the ant *Formica lugubris* (Walter et al. 1993). Although this function has not been shown in *S. invicta*, it could be that a similar pheromone is produced by virgin queens before the mating flight.

Several gene families show expression patterns consistent with subfunctionalisation

The 4 copies of the vitellogenin gene in *S. invicta* seem to be playing very different roles owing to their expression patterns. The equivalent gene for vitellogenin in *Drosophila melanogaster*, the yolk protein gene, is involved exclusively in egg formation and is expressed mostly in the fat bodies and oocytes (Isaac & Bownes 1982). Other social insects with the vitellogenin gene such as the honey bee *Apis mellifera* express it in queen fat bodies where it is related to egg production too (Guidugli-Lazzarini et al. 2008; Excels 1974). In the honey bee, however it has also been found to be expressed in workers, and to play different roles – ranging from brood care behaviour to immune response (Seehuus et al. 2007; Park et al. 2018; Amdam & Omholt 2003). In ants, the vitellogenin gene has expanded through several duplication events. In some ant species, members of these vitellogenin gene families appear to have caste-specific functions, in an example of subfunctionalisation (Morandin et al. 2014; Corona et al. 2013).

The results described here support and expand those previous findings in the vitellogenins of *S. invicta*. Out of the four vitellogenin genes identified in *S. invicta* (Wurm et al. 2011), two of them (Vg2 and Vg3) are exclusively expressed in queens, but they are expressed throughout all the tissues. This is not the pattern that would be expected if they were involved only in egg production, instead these expression patterns suggest that they may be playing a different role, for instance, as a signalling molecule. Vg1 is expressed in all castes, again, in most tissues, suggesting that its role is not to produce eggs, not least because egg production is limited to queens. This particular copy of the gene is particularly highly expressed in fat bodies in workers. Finally, Vg4 is again expressed in all castes, but only in nervous tissues, that is, antenna, brain and abdominal ganglia. This suggests a function completely different to egg production, and potentially related to communication or behaviour, especially given the additional functions of vitellogenin in other social insects such as honey bees.

The different odorant binding proteins (OBPs) in *S. invicta* also seem to play different roles based on their expression patterns. Based on these patterns, OBPs can be classified into three broad clusters, one of the groups of OBPs shows relatively high expression throughout all tissues, mostly in queens and workers (although some in males too). This group of OBPs includes Gp-9 OBP3, a gene which is strongly linked to the alternative social forms of *S. invicta* and that has been considered for decades as candidate gene for explaining the differences between these two phenotypes (Keller & Ross 1998; Lucas et al. 2015). This

group of OBPs not only cluster together by expression patterns, but also phylogenetically (Pracana et al. 2017), what is more, they are all found within the supergene region. The fact that all OBP genes in the supergene region show similar expression patterns suggests that they all may be playing a role in the differences in social organisation, not only OBP3. The fact that only this particular OBP has been proposed as a candidate and none of the others probably stems from the fact that OBP3 is very highly expressed (in some samples, the most highly expressed gene in the dataset). This would have made it easier to detect this gene and its association with social form in the original allozyme assays (Keller & Ross 1993). The OBPs in the other two expression pattern clusters are outside the supergene region. One of these groups shows a pattern which is in line with its main function in other arthropods – as proteins linked to chemical perception (reviewed in (Pelosi et al. 2005). They are expressed mostly in the antenna, but also in other tissues that may be related to chemical communication such as the brain, abdominal ganglia and external cuticular tissues. Ants use molecules in their cuticule such as cuticular hydrocarbons as a method to recognise each other (Hölldobler et al. 1990). These molecules are spread throughout the cuticle, and it is thus possible that some OBP genes participate in the production of such signals. In fact, in the wasp *Polistes dominulus*, OBPs have also been found in legs and wings (Calvello et al. 2003), adding evidence for some OBPs acting as chemical messengers (Pelosi et al. 2005). Finally, the third group of OBPs is have low expression in general, with occasional highly tissue and caste specific expression. These OBPs might perform very specific functions, and be activated only on specific occasions. For instance, OBPZ2 is expressed in the Dufour gland and the spermatheca of queens, whereas OBP9 is only expressed in male testes. This would suggest that these OBPs could have a role in sexual behaviour, in a similar way to OBPs found in the seminal fluid of *D.melanogaster* (Findlay et al. 2008).

Chemosensory proteins (CSPs) show patterns which are consistent with chemical communication. Most of them are expressed in the antenna and/or in cuticular tissues. They do vary, however, in caste-specificity, with some of them being more highly expressed in each caste. The patterns found in CSPs are those expected given their roles in other arthropods. For instance, they have also been found to be expressed across a wider variety of tissues in many arthropod systems including honeybee *Apis mellifera* embryos (Maleszka et al. 2007), the legs of the cockroach *Periplaneta americana* (Kitabayashi et al. 1998), the mouthparts of the moths *Helicoverpa armigera* and *H. assulta* (Liu et al. 2014) or gonadal tissues in the wing tissue of the locust *Locusta migratoria* (Ban et al. 2003). In *S. invicta* CSP1 was previously reported as being expressed in the antenna, with which the results shown here agree (González et al. 2009; Kulmuni et al. 2013). In all, our results support the idea of CSPs being involved in caste differences, potentially in the regulation, production and reception of caste-specific chemical signals.

Conclusion

Our results show that the expression differences between castes are highly tissue-specific. Most tissues have thousands of genes which are differentially expressed between castes, only a few hundred of genes are consistently caste biased across all tissues. These patterns support the idea that highly connected genes regulate the differences between castes in social insects. Some tissues are more specialised in terms of gene expression in specific castes, for instance, eyes have a larger proportion of male biased genes whereas antenna has more worker and queen biased genes. The gene families that are known to have expanded in the fire ant lineage and are likely to be involved in caste differences show expression patterns that are consistent with subfunctionalisation. More specifically, vitellogenins, OBPs and CSPs show tissue and caste specific patterns for which the functions may vary between egg production in queens to chemical communication.

Our results are the most extensive tissue-specific gene expression data for any hymenopteran generated to date. As such, they reveal new molecular mechanisms underlying one of the most spectacular polyphenisms in animals: morphological caste differences in social insects. Future work will expand on these results, by analysing other gene families of interest such as odorant receptors, and particular genes that are known to be involved in caste differentiation such as *doublesex*. Additionally, future work should also focus on reconciling genomic data with the transcription data provided here. For instance, testing whether duplicated genes are enriched in genes with caste biased expression would provide further evidence for the role of subfunctionalisation in caste differences. In addition, linking patterns of gene expression differences between castes with selection signatures in the genomic sequences would provide more depth into questions about the selective pressures emerging from caste differences.

References

Abouheif, E. & Wray, G.A., 2002. Evolution of the gene network underlying wing polyphenism in ants. *Science*, 297(5579), p.249–252.

Alexa, A. & Rahnenfuhrer, J., 2016. topGO: enrichment analysis for gene ontology. *R package version*.

Amdam, G.V. & Omholt, S.W., 2003. The hive bee to forager transition in honeybee colonies: the double repressor hypothesis. *Journal of theoretical biology*, 223(4), p.451–464.

Anderson, K.E., Linksvayer, T.A. & Smith, C.R., 2008. The causes and consequences of genetic caste determination in ants (Hymenoptera: Formicidae). *Myrmecological news / Osterreichische Gesellschaft fur Entomofaunistik*, 11, p.119–132.

Arrese, E.L. & Soulages, J.L., 2010. Insect fat body: energy, metabolism, and regulation. *Annual review of entomology*, 55, p.207–225.

Ban, L. et al., 2003. Chemosensory proteins of *Locusta migratoria*. *Insect molecular biology*, 12(2), p.125–134.

Berens, A.J., Hunt, J.H. & Toth, A.L., 2015. Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. *Molecular biology and evolution*, 32(3), p.690–703.

Beye, M. et al., 2003. The gene csd is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell*, 114(4), p.419–429.

Braendle, C. & Flatt, T., 2006. A role for genetic accommodation in evolution? *BioEssays: news and reviews in molecular, cellular and developmental biology*, 28(9), p.868–873.

Bray, N.L. et al., 2016. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5), p.525–527.

Buechel, S.D., Wurm, Y. & Keller, L., 2014. Social chromosome variants differentially affect queen determination and the survival of workers in the fire ant *Solenopsis invicta*. *Molecular ecology*, 23(20), p.5117–5127.

Calvello, M. et al., 2003. Soluble proteins of chemical communication in the social wasp *Polistes dominulus*. *Cellular and molecular life sciences: CMLS*, 60(9), p.1933–1943.

Chau, L.M. & Goodisman, M.A.D., 2017. Gene duplication and the evolution of phenotypic diversity in insect societies. *Evolution; international journal of organic evolution*, 71(12), p.2871–2884.

Corona, M. et al., 2013. Vitellogenin underwent subfunctionalization to acquire caste and behavioral specific expression in the harvester ant *Pogonomyrmex barbatus*. *PLoS genetics*, 9(8), p.1003730.

Dow, J.A.T., 2009. Insights into the Malpighian tubule from functional genomics. *The Journal of experimental biology*, 212(Pt 3), p.435–445.

Durinck, S. et al., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, 4(8), p.1184–1191.

Ehmer, B. & Gronenberg, W., 1997. Antennal muscles and fast antennal movements in ants. *Journal of comparative physiology. B, Biochemical, systemic, and environmental physiology*, 167(4), p.287–296.

Ewels, P. et al., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics,* 32(19), p.3047–3048.

Excels, W., 1974. Occurrence and significance of Vitellogenins in female castes of social Hymenoptera. *Integrative and comparative biology*, 14(4), p.1229–1237.

Feldmeyer, B., Elsner, D. & Foitzik, S., 2014. Gene expression patterns associated with caste and reproductive status in ants: worker-specific genes are more derived than queen-specific ones. *Molecular ecology*, 23(1), p.151–161.

Findlay, G.D. et al., 2008. Proteomics Reveals Novel *Drosophila* Seminal Fluid Proteins Transferred at Mating. *PLoS Biology*, 6(7), p.178.

Fjerdingstad, E.J. & Crozier, R.H., 2006. The evolution of worker caste diversity in social insects. *The American naturalist*, 167(3), p.390–400.

Foret, S. et al., 2012. DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proceedings of the National Academy of Sciences of the United States of America*, 109(13), p.4968–4973.

Gadagkar, R., 1997. The evolution of caste polymorphism in social insects: Genetic release followed by diversifying evolution. *Journal of genetics*, 76(3), p.167–179.

Gadau, J. et al., 2012. The genomic impact of 100 million years of social evolution in seven ant species. *Trends in genetics: TIG*, 28(1), p.14–21.

Galili, T., 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics,* 31(22), p.3718–3720.

Glastad, K.M. et al., 2016. The caste- and sex-specific DNA methylome of the termite *Zootermopsis nevadensis*. *Scientific reports*, 6, p.37110.

González, D. et al., 2009. The major antennal chemosensory protein of red imported fire ant workers. *Insect molecular biology*, 18(3), p.395–404.

Gremme, G., Steinbiss, S. & Kurtz, S., 2013. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *Transactions on computational biology and bioinformatics / IEEE, ACM*, 10(3), p.645–656.

Guidugli-Lazzarini, K.R. et al., 2008. Expression analysis of putative vitellogenin and lipophorin receptors in honey bee (*Apis mellifera* L.) queens and workers. *Journal of insect physiology*, 54(7), p.1138–1147.

Hamilton, N.E. & Ferry, M., 2018. ggtern: Ternary diagrams using ggplot2. *Journal of statistical software*, 87(1), p.1–17.

Heimpel, G.E. & de Boer, J.G., 2008. Sex determination in the Hymenoptera. *Annual review of entomology*, 53, p.209–230.

Helms, J.A., 2018. The flight ecology of ants (Hymenoptera: Formicidae). *Myrmecological news*, 16, p.19-30.

Hojo, M.K. et al., 2015. Antennal RNA-sequencing analysis reveals evolutionary aspects of chemosensory proteins in the carpenter ant, *Camponotus japonicus*. *Scientific reports*, 5, p.13541.

Hölldobler, B. et al., 1990. *The Ants*, Harvard University Press.

Howard, R.W. & Blomquist, G.J., 1982. Chemical ecology and biochemistry of insect hydrocarbons. *Annual review of entomology*, 27(1), p.149–172.

Huber, W. et al., 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, 12(2), p.115–121.

Isaac, P.G. & Bownes, M., 1982. Ovarian and fat-body vitellogenin synthesis in *Drosophila melanogaster*. *European journal of biochemistry / FEBS*, 123(3), p.527–534.

Jackson, B.D. & Morgan, E.D., 1993. Insect chemical communication: Pheromones and exocrine glands of ants. *Chemoecology*, 4(3), p.125–144.

Janzen, F.J. & Paukstis, G.L., 1991. Environmental sex determination in reptiles: ecology, evolution, and experimental design. *The Quarterly review of biology*, 66(2), p.149–179.

Jasper, W.C. et al., 2015. Large-scale coding sequence change underlies the evolution of postdevelopmental novelty in honey bees. *Molecular biology and evolution*, 32(2), p.334–346.

Johnson, B.R., Atallah, J. & Plachetzki, D.C., 2013. The importance of tissue specificity for RNA-seq: highlighting the errors of composite structure extractions. *BMC genomics*, 14, p.586.

Johnson, B.R. & Jasper, W.C., 2016. Complex patterns of differential expression in candidate master regulatory genes for social behavior in honey bees. *Behavioral ecology and sociobiology*, 70(7), p.1033–1043.

Keller, L. & Ross, K.G., 1993. Phenotypic plasticity and “cultural transmission” of alternative social organizations in the fire ant *Solenopsis invicta*. *Behavioral ecology and sociobiology*, 33(2), p.121–129.

Keller, L. & Ross, K.G., 1998. Selfish genes: a green beard in the red fire ant. *Nature*, 394, p.573.

Kitabayashi, A.N. et al., 1998. Molecular cloning of cDNA for p10, a novel protein that increases in the regenerating legs of *Periplaneta americana* (American cockroach). *Insect biochemistry and molecular biology*, 28(10), p.785–790.

Koch, S.I. et al., 2013. Caste-specific expression patterns of immune response and chemosensory related genes in the leaf-cutting ant, *Atta vollenweideri*. *PloS one*, 8(11), p.81518.

Krieger, M.J.B. & Ross, K.G., 2002. Identification of a major gene regulating complex social behavior. *Science*, 295(5553), p.328–332.

Kulmuni, J., Wurm, Y. & Pamilo, P., 2013. Comparative genomics of chemosensory protein genes reveals rapid evolution and positive selection in ant-specific duplicates. *Heredity*,

110(6), p.538–547.

Law, C.W. et al., 2018. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research*, 5, p.1408.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25(16), p.2078–2079.

Liu, Y.-L. et al., 2014. Unique function of a chemosensory protein in the proboscis of two *Helicoverpa* species. *The Journal of experimental biology*, 217(Pt 10), p.1821–1826.

Love, M.I., Wolfgang, H. & Simon, A., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12).

Lucas, C., Nicolas, M. & Keller, L., 2015. Expression of foraging and Gp-9 are associated with social organization in the fire ant *Solenopsis invicta*. *Insect molecular biology*, 24(1), p.93–104.

Maleszka, J. et al., 2007. RNAi-induced phenotypes suggest a novel role for a chemosensory protein CSP5 in the development of embryonic integument in the honeybee (*Apis mellifera*). *Development genes and evolution*, 217(3), p.189–196.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), p.10–12.

Mikheyev, A.S. & Linksvayer, T.A., 2015. Genes associated with ant social behavior show distinct transcriptional and evolutionary patterns. *eLife*, 4, p.04775.

Miura, T., 2005. Developmental regulation of caste-specific characters in social-insect polyphenism. *Evolution & development*, 7(2), p.122–129.

Morandin, C. et al., 2015. Caste-biases in gene expression are specific to developmental stage in the ant *Formica exsecta*. *Journal of evolutionary biology*, 28(9), p.1705–1718.

Morandin, C. et al., 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome biology*, 17, p.43.

Morandin, C. et al., 2014. Not only for egg yolk—functional and evolutionary insights from expression, selection, and structural analyses of Formica ant Vitellogenins. *Molecular biology and evolution*, 31(8), p.2181–2193.

Nakanishi, A. et al., 2009. Sex-specific antennal sensory system in the ant *Camponotus japonicus*: structure and distribution of sensilla on the flagellum. *Cell and Tissue Research*, 338(1), p.79–97.

Narendra, A. et al., 2011. Caste-specific visual adaptations to distinct daily activity schedules in Australian Myrmecia ants. *Proceedings of the Royal Society B: Biological Sciences*, 278(1709), p.1141–1149.

Okonechnikov, K., Conesa, A. & García-Alcalde, F., 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics,* 32(2), p.

292–294.

Ometto, L. et al., 2011. Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. *Molecular biology and evolution*, 28(4), p.1381–1392.

Oxley, P.R. et al., 2014. The genome of the clonal raider ant *Cerapachys biroi*. *Current biology: CB*, 24(4), p.451–458.

Park, H.G. et al., 2018. Honeybee (*Apis cerana*) vitellogenin acts as an antimicrobial and antioxidant agent in the body and venom. *Developmental and comparative immunology*, 85, p.51–60.

Pelosi, P., Calvello, M. & Ban, L., 2005. Diversity of odorant-binding proteins and chemosensory proteins in insects. *Chemical senses*, 30 Suppl 1, p.i291–2.

Phillips, J.E. et al., 1987. Mechanisms and control of reabsorption in insect hindgut. In P. D. Evans & V. B. Wigglesworth, eds. *Advances in Insect Physiology*. Academic Press, p. 329–422.

Pracana, R. et al., 2017. Fire ant social chromosomes: Differences in number, sequence and expression of odorant binding proteins. *Evolution letters*, 1(4), p.199–210.

Price, J. et al., 2018. Alternative splicing associated with phenotypic plasticity in the bumble bee *Bombus terrestris*. *Molecular ecology*, 27(4), p.1036–1043.

R Core Team, 2019. R: A language and environment for statistical computing.

Richard, F.-J. & Hunt, J.H., 2013. Intracolony chemical communication in social insects. *Insectes sociaux*, 60(3), p.275–291.

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7), p.47.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26(1), p.139–140.

Ross, K.G. & Keller, L., 1995. Joint Influence of Gene Flow and Selection on a Reproductively Important Genetic Polymorphism in the Fire Ant Solenopsis invicta. *The American naturalist*, 146(3), p.325–348.

Seehuus, S.-C. et al., 2007. Immunogold localization of vitellogenin in the ovaries, hypopharyngeal glands and head fat bodies of honeybee workers, *Apis mellifera*. *Journal of insect science,* 7, p.1–14.

Shik, J.Z., Donoso, D.A. & Kaspari, M., 2013. The life history continuum hypothesis links traits of male ants with life outside the nest. *Entomologia experimentalis et applicata*, 149(2), p.99–109.

Simpson, S.J., Sword, G.A. & Lo, N., 2011. Polyphenism in insects. *Current biology: CB*, 21(18), p.R738–49.

Slater, G.S.C. & Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6, p.31.

Soneson, C., Love, M.I. & Robinson, M.D., 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, p.1521.

Tschinkel, W.R., 1993. Resource allocation, brood production and cannibalism during colony founding in the fire ant,*Solenopsis invicta*. *Behavioral ecology and sociobiology*, 33(4), p.209–223.

Vander Meer, R.K., Alvarez, F. & Lofgren, C.S., 1988. Isolation of the trail recruitment pheromone of *Solenopsis invicta*. *Journal of chemical ecology*, 14(3), p.825–838.

Vargo, E.L., 1997. Poison gland of queen fire ants (*Solenopsis invicta*) is the source of a primer pheromone. *Die Naturwissenschaften*, 84(11), p.507–510.

Vargo, E.L. & Hulsey, C.D., 2000. Multiple glandular origins of queen pheromones in the fire ant *Solenopsis invicta*. *Journal of insect physiology*, 46(8), p.1151–1159.

Vu, V.Q., 2011. ggbiplot: A ggplot2 based biplot. R package version 0.55.

Walter, F. et al., 1993. Identification of the sex pheromone of an ant, *Formica lugubris* (Hymenoptera, Formicidae). *Die Naturwissenschaften*, 80(1), p.30–34.

Wheeler, D.E., 1991. The developmental basis of worker caste polymorphism in ants. *The American naturalist*, 138(5), p.1218–1238.

Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis*, Springer.

Wurm, Y. et al., 2011. The genome of the fire ant *Solenopsis invicta*. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14), p.5679–5684.

Zhang, W. et al., 2016. Tissue, developmental, and caste-specific expression of odorant binding proteins in a eusocial insect, the red imported fire ant, *Solenopsis invicta*. *Scientific reports*, 6, p.35452.

Chapter 5: Conclusion and future steps

Summary and final remarks

In this thesis I have explored the interplay between phenotypic differences and their molecular underpinnings. I have focused on two cases of complex stable polymorphisms where differences were either determined in the coding sequence or triggered by environmental factors. Both of these two types of phenotypic differences are present in the red fire ant *Solenopsis invicta*. In this social insect colonies can either have a single or multiple queens, a difference in social organisation that results in behavioural and physiological changes at the colony level. This difference in social form is determined by differences in the coding sequence, specifically, a supergene with two variants, SB and Sb. On the other hand, *S. invicta*, as many other ant species, has three morphologically distinct castes: queens, workers and males. In this case, all three castes share the same coding sequence. Through the analysis of expression patterns and modelling, I have investigated the impact of different evolutionary forces in interaction shaping the genome and transcriptome underlying complex phenotypic differences.

The supergene of *Solenopsis invicta* shows signs of resolved evolutionary conflict. This was shown by analysing the expression patterns between social forms and between variants of the supergene in two populations of *S. invicta* from North and South America. The results of this analysis showed that the supergene is enriched in genes with differential expression between social forms, with Sb (the variant present only in multiple-queen colonies) being enriched in multiple-queen biased genes. These patterns are consistent with those found in sex chromosomes in the context of sexual dimorphism, and suggest that the supergene is enriched with antagonistic loci. In other words, the supergene is enriched in genes which may have different fitness effects in either social form, and more specifically, Sb is enriched in genes which are likely to be beneficial for multiple-queen colonies. This suggests that evolutionary conflict could have resulted in reduced recombination between these antagonistic loci in the supergene region.

The results of the joint gene expression analyses also show that most of the expression differences between variants are unrelated to differences within social form. These variant-specific expression patterns could be due to the ongoing degeneration of Sb, as a result of the lack of recombination in this variant. For instance, most genes with expression differences between variants but no expression differences between social forms were more highly expressed in SB. SB does recombine normally, and is thus subjected to stronger purifying selection than Sb. This pattern of expression is consistent with gene specific dosage for genes with deleterious mutations in Sb. These results show that evolutionary

conflict is likely to play an important role in this particular supergene, but it does not give any information about the nature of such conflict.

Evolutionary conflict favouring the selection for supergenes can emerge from multiple processes, including from standing variation in antagonistic variants in two or more phenotypes in the same population or from locally adapted populations with gene flow between them. **A model using parameters from the life history of *S. invicta* showed that gene flow is the most likely process to favour the emergence of the social chromosome.** The model simulates the spread of antagonistic alleles with or without linkage to the supergene and under different levels of gene flow between social forms. The parameters controlling gene flow between social forms are estimated using data from natural populations. The results of the simulations show that, without linkage, antagonistic alleles tend to become fixed or lost in the population. There is only a narrow range of circumstances under which both are present in equilibrium in the population. Additionally, changes in the gene flow have a very strong effect in the outcome of the simulations. More specifically, the set of circumstances for which both antagonistic alleles exist at an equilibrium varies widely depending on the strength of the gene flow between social forms. In natural populations, and over evolutionary time, gene flow between social forms is likely to change, which implies that it is very unlikely that both antagonistic alleles exist at an equilibrium over an extended period of time. Consequently, it is very unlikely that selection would favour the formation of supergenes only from standing variation. When the antagonistic alleles were linked to the supergene, however, gene flow between social forms had a smaller impact on the spread of the antagonistic alleles. In addition, in most cases both alleles were kept in equilibrium in the population. From these results I hypothesised that the supergene emerged as a strategy to "shield" antagonistic alleles from the influence of gene flow, making it more difficult for any of the alleles to become lost in the population. Linkage to the supergene would not necessarily be the strategy that would maximise fitness at a given time point, but it would be the optimal strategy over evolutionary time.

The gene expression datasets available to date for ants in general, and *S. invicta* in particular, are not ideal for investigating differences between castes. This is because such datasets are often obtained from a few tissues or whole bodies, giving either a limited or an equivocal picture of gene expression differences between castes. For this thesis I generated a highly detailed tissue-specific gene expression dataset for adult workers, queens and males of *S. invicta*. This newly generated data shows that **gene expression differences between castes is highly tissue specific. Only a handful of genes are consistently differentially expressed between castes across most tissues.** This data also allowed a

more detailed analysis of specific gene families, such as OBPs, but also gene families encoding for vitellogenins and chemosensory proteins. The results confirm previous findings for these groups of genes, and builds upon them, giving a very detailed account of how tissue-specific patterns for these genes vary across caste. For instance, for OBPs we show that not only do they show caste-specific patterns, but also that these expression patterns correlate with their phylogeny. In other words, phylogenetically close OBPs are more likely to also have similar expression patterns. This includes the OBPs in the supergene, which form a monophyletic group. These results emphasize the idea that by expanding, some gene families acquire new functions and play different roles in different castes.

Overall, this thesis has explored a specific system of stable phenotypic differences and contextualised it in a wider theoretical framework. This works in two ways, firstly, I have successfully applied existing theory and tools from other systems to the very specific case of *S. invicta*. Secondly, the results obtained here expand and make a more general case for the existing theory on supergenes and complex stable polymorphisms. For instance, I have applied the concepts and techniques emerging from the study of sexual conflict and sex chromosomes to the social chromosome. This theoretical framework has allowed me to ask the relevant questions and to interpret the results in the light of years of previous research on a similar system. At the same time, it shows that the body of research reserved thus far to sex chromosomes and sexual conflict can be applied more broadly, to a whole set of increasingly relevant supergene systems. Similarly, the model I have built was based on previous work relating to linkage of sexually antagonistic alleles to sex chromosomes. The results I obtained are in line with predictions of earlier abstract theoretical models on the evolution of supergenes (e.g. Turner 1967). Therefore, by applying life history traits of *S. invicta* to existing models of sexual antagonism I have produced results which can be generalised using classic supergene theory. The results from this thesis are evidence that stable differences in complex phenotypes share many general evolutionary and molecular mechanisms. The body of work and theory based on one system can thus often be generalised to account for the wider phenomenon of discrete phenotypic differences in populations connected by gene flow. This could be of interest, for instance, in seemingly unrelated processes, such as speciation, where, evidence increasingly shows, supergenes may also play an important role (Kirkpatrick & Barton 2006; Barth et al. 2017; Lucek et al. 2019).

Finally, with the work carried out in this thesis I have generated the largest tissue-specific gene expression dataset for any hymenopteran to date. This dataset will allow future researchers to explore the molecular underpinning of caste differences with high resolution.

Additionally, this dataset will be integrated with already existing tissue-specific gene expression data in other organisms, allowing for the study of gene expression evolution across animals.

Of course, due to the time limitations of a PhD thesis, many interesting research questions cannot be addressed. In the following section I will address potential avenues of future research given the data and results obtained for this thesis.

Future work

Here I have shown that only a handful out of the >400 genes in the *S. invicta* supergene are potentially involved in differences between social forms. Using the techniques described here it is not possible to know with certainty which are these genes or what function they could be playing. Carrying out functional analyses in social insects is very challenging, the phenotypes of interest (colony behaviour) are complex, and the generation time is of about 3 years for *S. invicta*. Despite these difficulties, gene modifications have been recently carried out successfully in social insects (Yan et al. 2017; Trible et al. 2017). Here I have identified several candidate genes that could be functionally tested. Of special interest are the genes encoding for OBPs located in the supergene. Many of these have protein coding differences between variants, and at least one of them is present only in Sb. An assay where these genes were modified in Sb only (for instance, by methods such as CrISPR-Cas9) could provide detailed insight into which genes are relevant, and how, with regards to explaining the differences between social forms.

Many of the techniques used here to detect signs of conflict using gene expression patterns only do so indirectly. That is, general expression differences between phenotypes can give an idea of the extent of the evolutionary conflict, and can detect genes in which conflict has already been resolved (Mank 2017). None of the tools here can, therefore, detect genes currently under evolutionary conflict. To do so, it would be necessary to use similar approaches to those used in Cheng & Kirkpatrick (2016) or Wright et al. (2018), where measures of genetic diversity were used along with differential expression between sexes to detect ongoing sexual conflict. Theory goes that loci under sexual conflict will be under balancing selection in a population, as they spend 50% of the time in each sex (provided that they are on the autosomes, which is the case for most of them). Therefore, a harmful locus for one sex but beneficial for the other would simultaneously increase and decrease its frequency in a population with no sex bias. According to this idea, loci under conflict should show increased genetic diversity, as other loci under balancing selection do (e.g. immune

genes). This approach could be used in *S. invicta*, to detect genes under evolutionary conflict between social forms, but also between castes. For the comparison between social forms new data would need to be generated. For detecting genes under conflict between castes, however, future work could use existing genomic data, along with the tissue-specific expression data generated here. More generally, genomic data along with the expression differences between castes could be used to detect evolutionary patterns associated with caste differences. For instance, genes which are expressed in queens are expected to be under stronger selection than those expressed in workers. This is because queens transmit their genes directly to the next generations, whereas workers only do so indirectly through the queens to which they are related (Warner et al. 2017; Linksvayer & Wade 2016). This hypothesis could be tested with the data generated here. Similarly, selection against male genes could be even stronger. This idea comes from the fact that males also transmit their genes directly to the next generations but, in addition, they are haploid. Because haploid individuals are more likely to expose recessive alleles to selection, selection would be expected to also be stronger. To explore these ideas new genomic data would have to be generated. Taking colony as a biological replicate, we would ideally extract DNA from 50 colonies of each social form, to reach a similar sample size to that used in Cheng & Kirkpatrick (2016) or Wright et al. (2018). Within each colony, we would extract DNA from pools of about 10 individuals from each caste (workers, queens and males), amounting to a total of 300 DNA samples. Ideally, we would use colonies from the native South American range, to avoid any artefacts caused by the bottleneck of the invasive populations when measuring genetic diversity. At Queen Mary there are currently many individuals from several hundred colonies collected from the native range in 2016 and stored in freezers at -80ºC. This material could be used to perform this analysis, which would save the time and money of a field trip to collect new samples. Still, performing 300 DNA extractions and sequence that material will be time and resource- consuming. It would probably take a year and an additional grant to perform this next step.

Similarly, it would be interesting to investigate transposable element activity in different castes. According the superorganism hypothesis, in social insects, males and queens act as the germline of an individual, whereas workers represent the soma (Boomsma & Gawne 2018). Transposable elements are known to be mostly active in germline tissues in the majority of organisms (Bourque et al. 2018). According to the idea of superorganismality, transposable elements should therefore be active mostly in queens and males. Previous studies have shown that transposable elements are more highly expressed in reproductive individuals in other social insects (*e.g.* the honeybee; Wang et al., 2017). These studies, however, are based on expression data from whole bodies. With the data generated here,

we could fine tune which tissues in particular express transposable elements. Based on previous evidence (Bourque et al. 2018), we would expect the gonads of both males and queens to have higher expression of transposable elements than any other tissue. Interestingly, due to the differences in ploidy and life history between males and queens the patterns of transposable element expression are likely to be different in each caste. As discussed earlier, because males are haploid in ants, selection is likely to be stronger in this caste. Given than any deleterious effect of transposable element expression would be higher in males, selection against its expression in male gonads could be stronger than in that of queens. There are some transposable elements identified in *S. invicta* based on whole body expression data (*e.g.* Nipitwattanaphon et al., 2014). With the RNAseq dataset generated here, new transposable elements could be identified. To do so, the raw alignments could be used to find transposable elements from databases. Including manual curation, detecting and analysing new transposable elements in the fire ant should not take more than 6 months from now, and would be a very interesting next step for this project.

Finally, the model presented in this thesis has shown interesting qualitative general patterns. Namely, it shows that gene flow between social forms varies very strongly owing to demographic and environmental factors. It would be interesting to estimate, with a more quantitative approach, to what extent (and how) these different factors affect the gene flow. To this end, a bottom-up modelling approach such as an individual based model would be ideal and would produce highly relevant results to better understand the life history of *S. invicta*.

These future steps should build on the work described here, and expand our understanding of the complex interactions between genotypes and phenotypes. Ultimately, the body of work presented in this thesis and the new avenues of research it opens will result in a better understanding of the general processes governing the evolution of phenotypic diversity and plasticity.

References

Barth, J.M.I. et al., 2017. Genome architecture enables local adaptation of Atlantic cod despite high connectivity. *Molecular ecology*, 26(17), p.4452–4466.

Boomsma, J.J. & Gawne, R., 2018. Superorganismality and caste differentiation as points of no return: how the major evolutionary transitions were lost in translation. *Biological reviews of the Cambridge Philosophical Society*, 93(1), p.28–54.

Bourque, G. et al., 2018. Ten things you should know about transposable elements. *Genome biology*, 19(1), p.199.

Cheng, C. & Kirkpatrick, M., 2016. Sex-specific selection and sex-biased gene expression in humans and flies. *PLoS genetics*, 12(9), p.1006170.

Kirkpatrick, M. & Barton, N., 2006. Chromosome inversions, local adaptation and speciation. *Genetics*, 173(1), p.419–434.

Linksvayer, T.A. & Wade, M.J., 2016. Theoretical predictions for sociogenomic data: the effects of kin selection and sex-limited expression on the evolution of social insect genomes. *Frontiers in Ecology and Evolution*, 4, p.E1.

Lucek, K., Gompert, Z. & Nosil, P., 2019. The role of structural genomic variants in population differentiation and ecotype formation in *Timema cristinae* walking sticks. *Molecular ecology*, 28(6), p.1224–1237.

Mank, J.E., 2017. The transcriptional architecture of phenotypic dimorphism. *Nature ecology & evolution*, 1(1), p.6.

Trible, W. et al., 2017. orco mutagenesis causes loss of antennal lobe glomeruli and impaired social behavior in ants. *Cell*, 170(4), p.727–735.e10.

Turner, J.R.G., 1967. On supergenes. I. The evolution of supergenes. *The American naturalist*, 101(919), p.195–221.

Wang, W. et al., 2017. Contrasting sex-and caste-dependent piRNA profiles in the transposon depleted haplodiploid honeybee *Apis mellifera*. *Genome biology and evolution*, *9*(5), 1341-1356.

Warner, M.R., Mikheyev, A.S. & Linksvayer, T.A., 2017. Genomic signature of kin selection in an ant with obligately sterile workers. *Molecular biology and evolution*, 34(7), p.1780–

1787.

Wright, A.E. et al., 2018. Male-biased gene expression resolves sexual conflict through the evolution of sex-specific genetic architecture. *Evolution Letters*, 2(2), p.52–61.

Yan, H. et al., 2017. An engineered orco mutation produces aberrant social behavior and defective neural development in ants. *Cell*, 170(4), p.736–747.e9.

Annex I

Texts

Text AI.1: RNA extraction protocol

Preparation:

1P) Label as many screw-cap 2mL tubes as samples to extract.

2P) Label twice as many 1.5 mL Eppendorf tubes as samples to extract (one batch for RNA, the other for DNA)

3P) Fill tubes with 1g of ceramic beads and keep the tubes in dry ice.

4P) Move samples from the -80ºC freezer into dry ice.

5P) Use soft and hard forceps in a petri dish over dry ice to separate the abdomen, the thorax and the head of the queens as well as the the head of the workers.

6P) Move all parts of queens into separate 2mL screw-cap tubes.

7P) Move the body of the worker into a screw-cap tube. The head will be kept in the -80ºC freezer in case the DNA extraction with Trizol fails.

8P) Add 400 µL of Trizol to all samples, still in dry ice.

9P) Homogenise samples in the FastPrep 1800 rpm 1 min, then 1 min to settle, 3x

10P) Add 80 µL chloroform, vortex 10 sec, save aqueous for RNA. Incubate at room temperature for 3 min.

11P) Pipette aqueous phase to the RNA tubes (Use RNAse-free tubes). Save pink (organic) phase for DNA extraction. See below for DNA protocol

RNA extraction

1R) Pipette 200 µL of isopropanol (RNAse-free) into RNA tubes.

2R) Centrifuge at maximum speed 30 minutes, at 4ºC.

3R) Wait overnight at -20 ºC, the RNA “drops” to the bottom and forms a better, but still small, pellet.

4R) Wash with 75% ethanol 100 µL, 1-5 min RT, centrifuge maximum speed 5 min

6R) Air dry upside-down on a towel. Dissolve in 11 µL of RNAse-free water.

8R) Qiagen RNeasy microkit spin column for RNA purification

9R) Store at -80 C.

DNA extraction :

1D) From 11P, transfer pink organic phase to a new tube to separate it from the beads.
2D) Add 120 µL of 100% ethanol to the pink organic phase. Mix well by flicking, do not vortex. It can store the DNA at 4 C overnight at this step.
3D) Wait 5 min, centrifuge at 2000 g for 5 minutes at 4º C. Some ant debris may remain, this is not a problem.
4D) Carefully remove the liquid using a pipette and discard. Leave any solid material.
5D) Wash DNA with 400 µL of 10% ethanol, 0.1M NaCitrate, mix, wait 15 min, mix, wait 15 min (or overnight).
6D) Centrifuge 2000 g for 5 minutes at 4º C. Remove liquid by pipetting.
7D) Repeat 5D + 6D
7D) Wash DNA with 800 µL of 75% ethanol, mix, wait 15 min, mix, wait 15 min
8D) Centrifuge 2000 g for 5 min and carefully remove liquid. Invert to dry on a towel (~15 minutes).
7D) Dissolve DNA in 38-40 µL of fresh 8mM NaOH. Leave overnight at 4 Cº.
8D) Centrifuge at maximum speed for 10-15 minutes.
9D) Transfer liquid to new tubes. Add 2 µL of 1M Tris pH 7.6
10D) GenElute Mammalian Genomic DNA Miniprep Kit spin column for DNA cleaning.
11D) Store at -20º C.

Text AI.2: Quality control of RNAseq datasets and alignment to reference

For all datasets, we assessed read quality using fastQC (v0.11.5; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We removed low quality bases using fqtrim with default parameters (v0.9.5; http://ccb.jhu.edu/software/fqtrim/) and Illumina adapters using Cutadapt (v1.13; Martin 2011). We then generated a STAR (v2.5.3a; Dobin et al. 2013) index of the *S. invicta* reference genome (version gnG; RefSeq GCF_00018807; (Wurm et al. 2011) while providing geneset v000188075.1 in GFF format through the 'sjdbGTFtagExonParentTranscript=Parent' option. As recommended by STAR's developers, we aligned each sample to the reference twice, using the 'out.tab' file for the second run, and set 'sjdbOverhang' to the maximum trimmed read length minus one, *i.e.,* 74 for the Wurm et al. (2011) and (Morandin et al. 2016) data and 149 for the South American colonies. Alignments were run in parallel using GNU Parallel (v20150922; Tange 2011). All these steps and downstream analyses were performed on the Queen Mary University of London's High Performance Cluster (King, Butcher, and Zalewski 2017).

We further assessed aligned reads (i.e., BAM files) using MultiQC (v1.5; Ewels et al. 2016) (results available in Supplementary Material) and the BodyGene_coverage.py script of the RSeQC toolkit (v2.6.4; Wang et al. 2016).

Text AI.3: South American populations DNA-seq details

For each individual, we extracted 1 µg of genomic DNA using a phenol-chloroform protocol. The extracted material was sheared to 350 bp fragments using a Covaris (M220). We constructed individually barcoded libraries using the Illumina TruSeq PCR-free kit. The libraries were quantified through qPCR (NEB library quant kit). An equimolar pool of the 46 libraries was sequenced on a HiSeq4000 at 150bp paired reads. The sequencing produced an average of 17,790,416 pairs of reads per sample, with a maximum of 38,823,285 and a minimum of 7,910,042.

Text AI.4: Additional steps for the GATK analysis

A reading group ID per sample was added to all the generated BAM files using the 'AddOrReplaceReadGroups' tool from Picard (v 2.7.0-SNAPSHOT; http://broadinstitute.github.io/picard/). GATK needs a genome reference index and dictionary ('fai' and a 'dict' files) to work. The former was generated using the 'index' option from Samtools (v1.3.1; Li et al. 2009) and the latter using the tool 'CreateSequenceDictionary' from Picard.

Text AI.5: Individual ASE analyses for the North and South American populations

The Wurm et al. (2011) RNA-seq data was analysed using 'sample' as a blocking factor and 'variant effect' as variable of interest. The South American RNA-seq data was analysed using 'body part' and 'colony of origin' as blocking factors, and the 'variant effect' as variable of interest. This analysis allowed us to detect differences in expression between variants specific to body part. Preliminary analyses showed that the interaction between 'variant effect' and 'body part' had no significant effect in any of the genes, and consequently, only the main 'variant effect' was considered as the factor of interest for this analysis.

In both analyses, the gene differences between variants are reported as logarithmic fold differences between the SB and the Sb counts. That is, genes with biased expression towards SB will produce positive logarithmic fold differences whereas those biased towards

Sb will produce a negative value. To check whether there was an overall bias towards either variant, we tested the significance of the deviation from 0 for the median logarithmic fold differences between SB and Sb via a Wilcoxon sum rank test.

Text AI.6: Results for within-population supergene variant specific expression analyses

69 genes within the supergene region were considered for the analysis of the South American dataset. From these, 10 were differentially expressed between variants, 6 being more highly expressed in SB and 4 in Sb (Annex I, Table AI.4a). Overall, the variant-specific expression levels were similar between SB and Sb (Median of logarithm fold differences = -0.06, Wilcoxon sum rank test p value = 0.38). From all differentially expressed genes between variants in South America, 50% (i.e. 5 out of 10) were also differentially expressed in North America according to the linear mixed effect model (Fig. 2.1).

This analysis was repeated using queens from North American populations of *S. invicta* Wurm et al. (2011) (Annex I, Fig AI.2, Table AI.4b). Out of these, 27 showed significant allele-specific expression. Of all variant-specific differentially expressed genes, 15 were more highly expressed in SB, and 12 in Sb, a non-significant difference from what would be expected by chance (Binomial test p value = 0.7). As with the populations from the native range, we found no significant bias in expression of either variant (Median of logarithm fold differences = -3×10^{-17}, Wilcoxon sum rank test p value = 0.45).

The independent analysis in each population has less power than the joint analysis with the linear mixed effect models, because the information across populations is lost. In spite of this, the independent gene expression analyses found 2 genes that were significantly differentially expressed across populations that were also detected in the linear mixed effect model, namely: ‘ejaculatory bulb-specific protein 3’ (LOC105193134) and ‘carbohydrate sulfotransferase 11-like’ (LOC105199531).

Text AI.7: Expression differences between social forms

To add more robustness to the results obtained from the comparison between social forms (Morandin et al. (2016) data), the expression levels of single-queen queens from the Morandin et al. (2016) data were compared to those of the multiple-queen queens from the Wurm et al. (2011) data. For this, the samples were normalised together using the default DESeq2 method. Because these samples belong to different datasets, there are too many

confounding factors to estimate significance levels of differential expression. Instead, a default DESeq2 analysis was carried out to obtain the logarithm2 fold differences of this comparison. These log2 fold differences were then compared using a linear regression to those obtained from the comparison using only the (Morandin et al. 2016) data to detect common patterns in both datasets. All statistical tests were carried out using R (v3.4.4; R Core Team 2017).

Tables

Table AI.1: Detailed information for the RNA-seq datasets used in this study. a) Accession numbers of the North American RNA-seq datasets. “Project” and “SRA” columns indicate NCBI identifiers. The descriptions provided and the sequencing method used are based on metadata available on NCBI and in the manuscripts. A sample (marked with an asterisk) was discarded because of very low coverage after aligning the reads to the *S. invicta* genome. b) Details for the South American RNAseq dataset. From left to right, the colony name from where samples were taken, the caste used from these colonies, the boy parts extracted, the location of each colony in Argentina, the coordinates from where the sample was taken and finally, whether or not samples from the same colony were used to generate the VCF with fixed differences between Sb and SB. The samples marked with asterisks were discarded because of very low coverage after aligning the reads to the S. invicta genome. Samples marked with a cross were not used in downstream analyses to keep a balanced design.

a)

Publication	Project	Sequencing method	Body part	SRA	Description
Morandin et al 2016	**PRJDB4088**	**Illumina paired-end**	Whole body	DRS023315	Pool of 3 polygyne queen 1
			Whole body	DRS023316*	Pool of 3 polygyne queen 2
			Whole body	DRS023317	Pool of 3 polygyne queen 3
			Whole body	DRS023309	Pool of 3 monogyne queen 1
			Whole body	DRS023310	Pool of 3 monogyne queen 2
			Whole body	DRS023311	Pool of 3 monogyne queen 3
Wurm et al 2011	**PRJNA49629**	**Illumina single-end**	Whole body	SRS377035	4 pooled polygyne queens 1
			Whole body	SRS376911	4 pooled polygyne queens 2
			Whole body	SRS376910	4 pooled polygyne queens 3
			Whole body	SRS376905	4 pooled polygyne queens 4
			Whole body	SRS376904	4 pooled polygyne queens 5
			Whole body	SRS376903	4 pooled polygyne queens 6

b)

Colony	Caste	Body part	Town	Coordinates	Represented in VCF?
AR43	Female alate	Thorax, head, abdomen	Ceres	-29.88, -61.94	No
	Worker	Whole body			
AR3	Female alate	Thorax, head, abdomen	Alta Gracia	-31.65, -64.44	No
	Worker	Whole body			
AR28†	Whole body	Worker	Tanti	-31.37, -64.52	Yes
AR118	Female alate	Thorax, head, abdomen	San Carlos	-27.75, -55.89	No
	Worker	Whole body			
AR117*	Female alate	Thorax, head, abdomen	Vellejos Cué	-27.61, -57.00	No
	Worker	Whole body			
AR114†	Worker	Whole body	Berón de Astrada	-27.53, -57.56	Yes
	Female alate	Thorax, head, abdomen			
AR112	Female alate	Thorax, head, abdomen	Berón de Astrada	-27.53, -57.56	Yes
	Worker	Whole body			
AR111	Female alate	Thorax, head, abdomen	Berón de Astrada	-27.53, -57.56	Yes
AR104	Whole body	Worker	Barranqueras	-27.46, -58.91	No

Table AI.2: Location of the colonies used to estimate single nucleotide polymorphisms (SNPs) between SB and Sb males. Note that individuals from colonies AR102, AR111, AR112, AR114 and AR28 were also used to extract RNA for the RNAseq dataset.

Colony	Town	Coordinates
AR102	Barranqueras	-27.46, -58.91
AR103	Barranqueras	-27.46, -58.91
AR111	Berón de Astrada	-27.53, -57.56
AR112	Berón de Astrada	-27.53, -57.56
AR113	Berón de Astrada	-27.53, -57.56
AR114	Berón de Astrada	-27.53, -57.56
AR116	Vellejos Cué	-27.61, -57.00
AR122	Carlos Pellegrini	-28.54, -57.18
AR179	Villa Maria	-32.39, -63.24
AR181	Villa Maria	-32.39, -63.24
AR28	Tanti	-31.37, -64.52
AR46	Santa Fe - Santa Rosa	-31.56, -60.52
AR58	Médanos	-33.44, -59.07

Table AI.3: Table showing the name and RefSeq ID of the seven genes that are significantly differentially expressed between the SB and Sb variants of the *S. invicta* supergene. The significance levels were determined using a linear mixed effect models on the logarithm2 fold change of the expression differences between SB and Sb. Population was used as a random effect and the logarithm2 fold changes were weighted by read count of the gene. The third column in the table shows whether that particular gene is also differentially expressed in the comparison between social forms (using (Morandin et al. 2016) data), and if so, in which social form it is more highly expressed.

Gene description	Gene RefSeq ID	Differentially expressed in social forms
carbohydrate sulfotransferase 11-like	LOC105193134	No
uncharacterized LOC105193135	LOC105193135	No
pheromone-binding protein Gp-9	LOC105194481	Higher in multiple-queen
retinol-binding protein pinta-like	LOC105199327	Higher in mutiple-queen
ejaculatory bulb-specific protein 3	LOC105199531	Higher in multiple-queen
uncharacterized LOC105199756	LOC105199756	No
calcium-independent phospholipase A2-gamma	LOC105203065	No
NADH dehydrogenase	LOC105199789	Higher in multiple-queen
Tankyrase	LOC105193132	No

Table AI.4: Genes with significant differential expression between the SB and Sb variant of the *S. invicta* supergene in **a)** South American or **b)** North American populations. The significance levels were determined by the Wald test in DESeq2. Significance was established as Benjamini and Hochberg corrected p values <0.05. The columns in the tables show the name of the gene, its ID number in RefSeq, their logarithm2 fold difference for the comparison between variants (values > 0 are more highly expressed in SB) and in which variant they are more highly expressed

a)

Gene description	Gene RefSeq ID	Log2 fold difference	Highly expressed in
tankyrase	LOC105193132	0.95	SB
carbohydrate sulfotransferase 11-like	LOC105193134	1.44	SB
uncharacterized LOC105193135	LOC105193135	-1.02	Sb
ejaculatory bulb-specific protein 3	LOC105199531	5.74	SB
NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial-like	LOC105199789	-3.95	Sb
ubiquitin-like domain-containing CTD phosphatase 1	LOC105199797	-1.10	Sb
cytochrome P450 4C1	LOC105202818	1.35	SB
uncharacterized LOC105203040	LOC105203040	-1.01	Sb
acyl-CoA Delta(11) desaturase	LOC105204968	2.57	SB
suppressor of lurcher protein 1	LOC105204975	1.60	SB

b)

Gene description	Gene RefSeq ID	Log2 fold difference	Highly expressed in
carbohydrate sulfotransferase 11-like	LOC105193134	1.30	SB
leucine-rich repeat-containing protein 15-like	LOC105193138	2.43	SB
uncharacterized LOC105193825	LOC105193825	-1.87	Sb
histone RNA hairpin-binding protein	LOC105193831	0.91	SB
modular serine protease-like	LOC105194434	-2.48	Sb
pheromone-binding protein Gp-9	LOC105194481	-0.76	Sb
pheromone-binding protein Gp-9-like	LOC105194501	-1.68	Sb
NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial-like	LOC105194902	-2.39	Sb

nanos homolog 3-like	LOC105195826	-1.74	Sb
alpha-2-macroglobulin receptor-associated protein	LOC105199295	-0.88	Sb
dynein regulatory complex subunit 4	LOC105199315	-2.57	Sb
galactoside 2-alpha-L-fucosyltransferase 2-like	LOC105199316	-1.14	Sb
retinol-binding protein pinta-like	LOC105199327	-1.52	Sb
retinol-binding protein pinta-like	LOC105199330	-1.60	Sb
uncharacterized LOC105199332	LOC105199332	1.73	SB
ejaculatory bulb-specific protein 3	LOC105199531	1.34	SB
uncharacterized LOC105199756	LOC105199756	3.03	SB
cytochrome P450 4C1	LOC105199861	-2.33	Sb
pre-mRNA-splicing factor Slu7	LOC105202812	4.54	SB
uncharacterized LOC105202816	LOC105202816	2.88	SB
pheromone-binding protein Gp-9	LOC105202823	2.94	SB
calcium-independent phospholipase A2-gamma	LOC105203065	1.98	SB
uncharacterized LOC105203078	LOC105203078	2.38	SB
cytochrome P450 4C1-like	LOC105203086	-2.94	Sb
histone-binding protein N1/N2-like	LOC105203091	0.59	SB
uncharacterized LOC105204969	LOC105204969	1.77	SB
uncharacterized LOC105207492	LOC105207492	1.19	SB

Figures

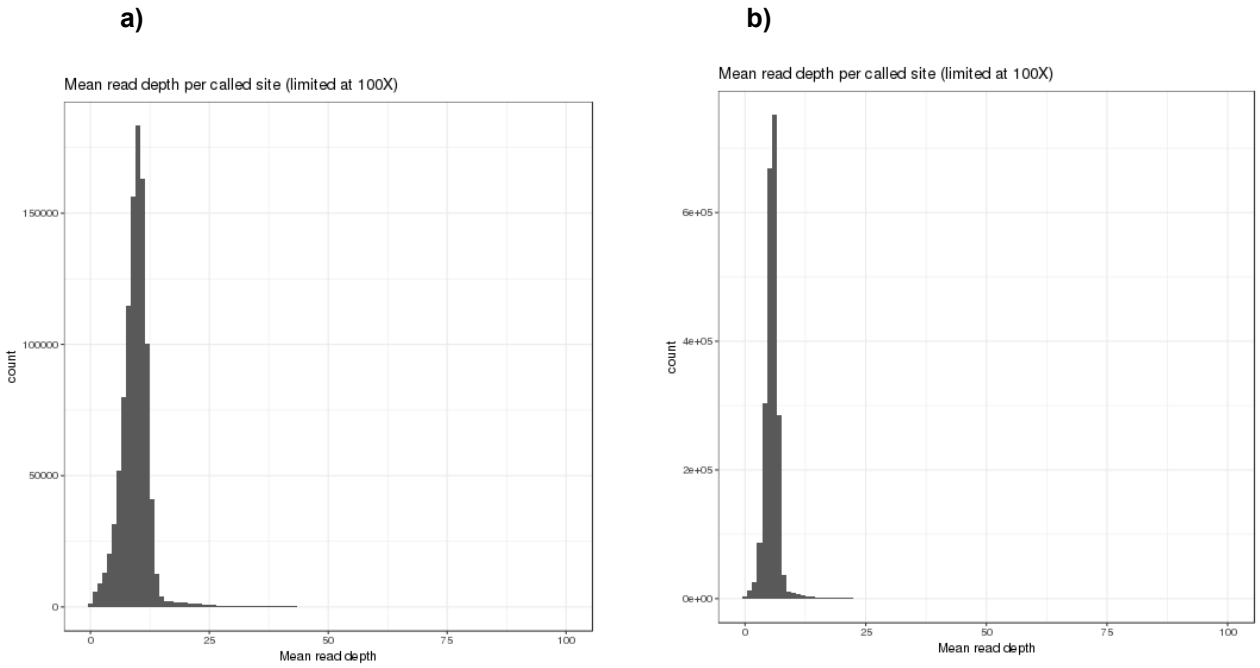
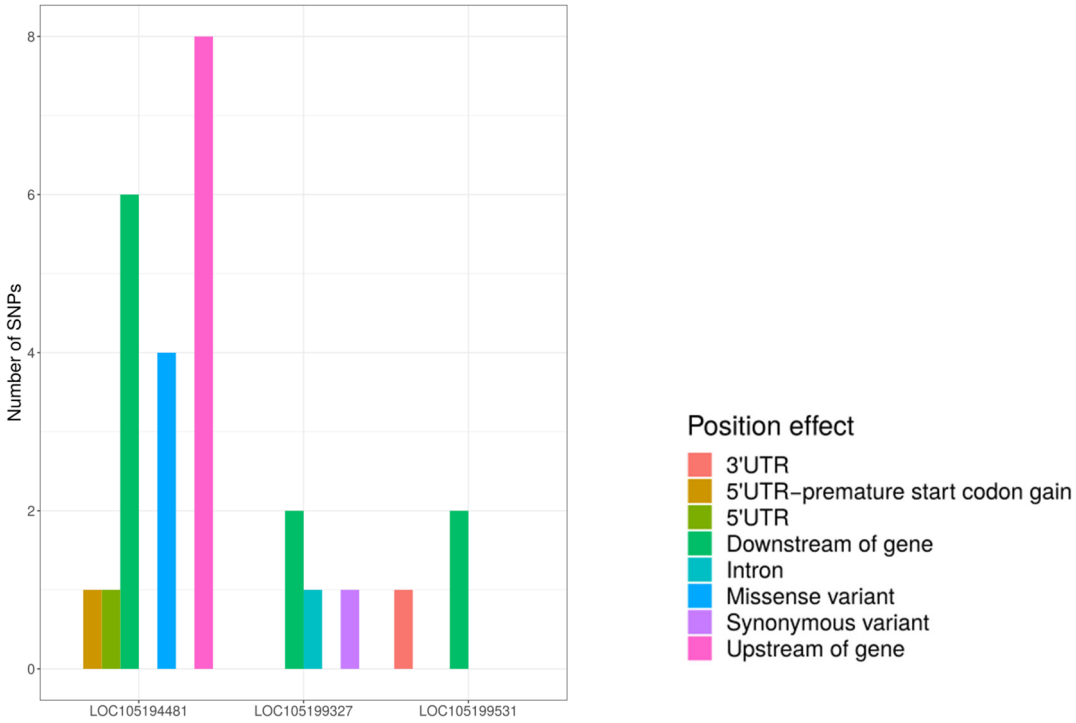


Figure AI.1: Frequency of mean read depth per called single nucleotide position (SNP) for **a)** North American populations and **b)** South American populations. Any SNPs with mean coverage higher than 16 for the North American samples and 12 for the South American samples were discarded from downstream analyses.

a)



b)

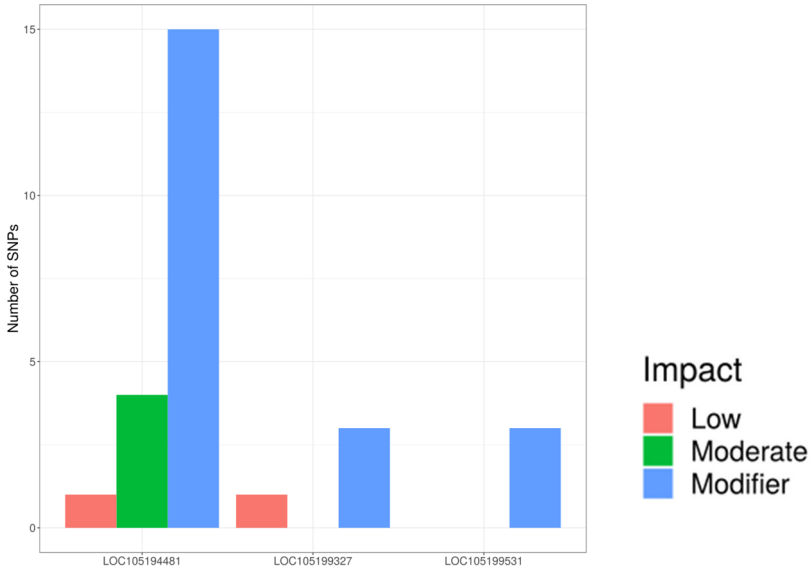


Figure AI.2: Effect of single nucleotide polymorphisms (SNPs) in Sb on the three candidate genes as reported by SNPeff. **a)** Shows the number of SNPs that affect different positions within the gene, **b)** shows the number of SNPs depending on their potential impact in the protein coding sequence.

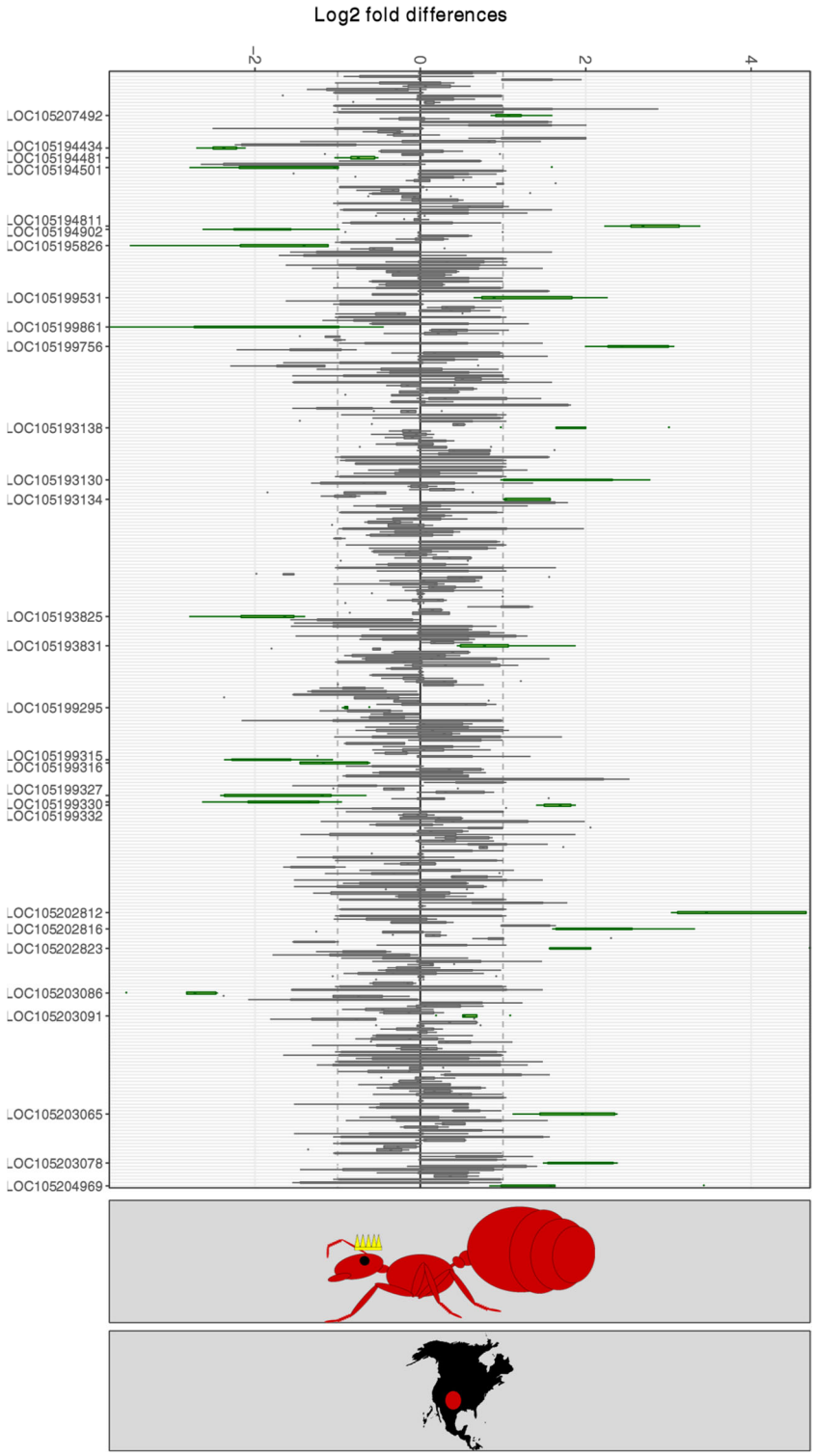


Figure AI.3: Allele-specific expression for genes in the fire ant social supergene. Differences in expression (y axis) between the variants of the social chromosome North American queens only (whole body). Only relative expression levels for social chromosome genes that had fixed SNP differences between SB and Sb are shown. Significantly expression differences; are indicated by green boxes (BH adjusted p value < 0.05 from Wald test in DESeq2). Non-significant differences are marked by grey boxes. Within each plot, each box shows the distribution of logarithm 2 fold differences between SB and Sb per caste/body part/population. Genes with values above the solid line (logarithm 2 fold differences = 0) are more highly expressed in SB and vice versa.

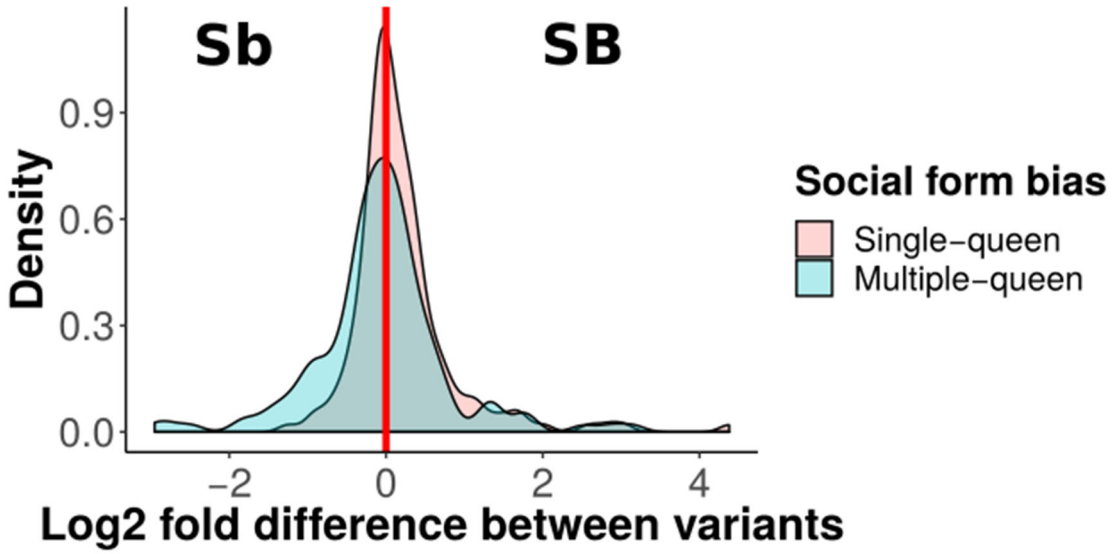


Figure AI.4: Density distribution of the logarithm2 differences for the comparison between variants of the supergene. The plot shows two distributions, corresponding to genes biased towards single-queens (pink) and for genes biased towards multiple queens (blue). Genes with negative log2 fold differences between variants are biased towards Sb, genes with positive log2 fold differences are biased towards SB. Note that the distribution for multiple-queen biased genes is more biased towards Sb than that of single-queen biased genes. This difference in the distribution is significant (KS test p value < 0.05). The differences between social forms were calculated using data from Morandin et al. (2016) and those between variants using data from Wurm et al. (2011). The distribution curves are based on data from 314 genes in the supergene region. These correspond with genes for which there was expression data for both comparisons.

References

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* 29 (1): 15–21.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32 (19): 3047–48.

King, T., S. Butcher, and L. Zalewski. 2017. Apocrita—High Performance Computing Cluster for Queen Mary. University of London.

Li, Heng, et al.,. 2009. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25 (16): 2078–79.

Martin, Marcel. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17 (1): 10–12.

Morandin, Claire, et al. 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome Biology* 17 (March): 43.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2017. https://www.R-project.org/.

Tange, O. 2011. Gnu Parallel-the Command-Line Power Tool. *The USENIX Magazine* 36 (1): 42–47.

Wang, Liguo, et al. 2016. Measure Transcript Integrity Using RNA-Seq Data. *BMC Bioinformatics* 17 (February): 58.

Wurm, Yannick, et al. 2011. “The Genome of the Fire Ant Solenopsis Invicta.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (14): 5679–84.

Annex II

Texts

Text AII.1: DNA Phenol-Chloroform Extraction for *Solenopsis invicta* workers

1. Sample preparation

- Whole ant in an eppendorf tube with 1.5 ml deionised water. 1 hour at RT (room temperature)
- Meanwhile, make up a sodium acetate-ethanol solution. Add 40mL 100% ethanol and 2mL 3M sodium acetate solution.

2. Homogenisation

- Set the heat block at 55 degrees C.
- In an eppendorf, add 350 ul CTAB, 10ul Proteinase K (10 mg/mL).
- Transfer the ant from the water eppendorf to a clean tissue to remove as much water as possible, then put it in the CTAB + Prot K eppendorf. Use a pestle to reduce the ant into small fragments.
- Incubate the ant solution on the heat block for an hour.

3. Phase separation (in fumehood)

- Pipette 350ul of the lower phase of phenol in each sample tube. Close lid, shake lightly 10 times. Load the samples in the centrifuge and leave to rest for 10 min. Centrifuge 10 minutes, full power
- Meanwhile, label 2 series of eppendorf tubes and add 350ul of chloroform in each.
- Label a 3rd series of eppendorf tubes and add 1mL sodium acetate-ethanol solution (prepared earlier). Put series 1 and 2 on the bench near centrifuge, and series 3 in the freezer.
- Pipette the upper phase of each sample into series 1 of eppendorf. Close the lid, shake lightly 10 times. Load the samples in the centrifuge and leave to rest for 5 min. Centrifuge 10 minutes, full power.
- Pipette the upper phase into series 2 of eppendorf. Close the lid, shake lightly 10 times. Load the samples in the centrifuge and run for 5 minutes, full power.
- Pipette the supernatant into series 3 of eppendorf. Expect to see a DNA pellet. Store in -20 freezer for an overnight precipitation.

4. DNA precipitation

- Centrifuge at full speed for 20 min. A DNA pellet should be visible.
- Remove as much supernatant as possible.
- Add 500ul of 70% ethanol. Close lid, swing twice.
- Centrifuge at full speed for 20min.
- Remove as much supernatant as possible.
- Add 500ul of 70% ethanol. Close lid, swing twice.
- Centrifuge at full speed for 20min.
- Remove all supernatant. Air dry for 10 minutes at RT.
- Once no ethanol remains in the tube, dissolve the DNA pellet in 20 - 30 ul of low TE buffer or deionised water.

Text AII.2: RNA extraction protocol for *S. invicta* tissues

The samples already have 200uL of Trizol, stored in -80ºC freezer

1.Homogenisation

- 1 minute at full speed (1800 rpm) in the FastPrep96
- Add 800uL Trizol and incubate 5 min at RT

2. Phase separation (in fumehood)

- Add 0.2mL of chloroform
- Vortex for 10 seconds
- Incubate 3 min at RT
- Centrifuge at 12000g for 15 min at 4ºC
- Transfer aqueous phase to a fresh nuclease-free 1-5 mL tube

3. RNA precipitation

- Add 0.5mL of 100% isopropanol
- Vortex the solution
- Incubate at RT for 20 min
- Centrifuge at 12000g for 10 min at 4ºC
- Leave sample overnight at -20ºC
- Centrifuge at 12000g for 5 min at 4ºC
- Remove the supernatant from the tube

4. RNA wash

- Wash the pellet with 1mL of 75% ethanol
- Vortex briefly, centrifuge at 7500 g for 5 min at 4ºC

- Discard the wash
- Airdry RNA pellet for 5 min
- Re-suspend in 50uL of nuclease-free water
- To RNAeasy kit cleaning.

5. RNeasy Mini Kit (Qiagen) spin column RNA cleanup step

- Adjust the sample to a volume of 100 µl with RNase-free water. Add 350 µl Buffer RLT, and mix well.
- Add 250 µl ethanol (96–100%) to the diluted RNA, and mix well by pipetting. Do not centrifuge
- Transfer the sample (700 µl) to an RNeasy Mini spin column placed in a 2 ml collection tube. Close the lid. Centrifuge for 15 s at ≥8000 x g. Discard the flow-through.
- Add 500 µl Buffer RPE to the RNeasy spin column. Close the lid. Centrifuge for 15 s at ≥8000 x g to wash the membrane. Discard the flow-through.
- Add 500 µl Buffer RPE to the RNeasy spin column. Close the lid. Centrifuge for 2 min at ≥8000 x g to wash the membrane.
- Place the RNeasy spin column in a new 2 ml collection tube. Close the lid, and centrifuge at full speed for 1 min.
- Place the RNeasy spin column in a new 1.5 ml collection tube. Add 30–50 µl RNase-free water directly to the spin column membrane. Close the lid, and centrifuge for 1 min at ≥8000 x g to elute the RNA.

Tables

Table AII.1. Complementary information about the colonies from which individuals were sampled for RNA extraction. Columns, from right to left: Colony ID, location of the colony (latitude), location of the colony (longitude), caste used for RNA extraction and order in which the individuals from each colonies were taken and snap frozen, date in which the dissection of the pool of 3 individuals per caste and colony was performed (colony 7 had a

Colony	GPS coordinates latitude	GPS coordinates longitude	Caste extracted	Sampling order	Dissection date
1	30.308863 N	97.591755 W	Males	1	10/01/2018
5	30.308668 N	97.591361 W	Reproductive females	8	10/01/2018
3	30.308795 N	97.591650 W	Workers	4	11/01/2018
2	30.308786 N	97.591629 W	Workers	3	15/01/2018
6	30.308623 N	97.591288 W	Males	10	15/01/2018
19	30.518678 N	96.424410 W	Reproductive females	13	16/01/2018
15	30.518455 N	96.424608 W	Males	9	18/01/2018
9	NA	NA	Workers	14	05/02/2018
11	30.518821 N	96.424913 W	Reproductive females	2	05/02/2018
8	30.308511 N	97.591013 W	Reproductive females	12	06/02/2018
13	30.518556 N	96.424056 W	Males	5	09/02/2018
4	30.308765 N	97.591350 W	Workers	6	13/02/2018
14	30.518533 N	96.424538 W	Workers	7	21/02/2018
7	30.308596 N	97.591175 W	Reproductive females	11	05/03/2018
20	30.518663 N	96.424358 W	Males	15	06/03/2018

Figures

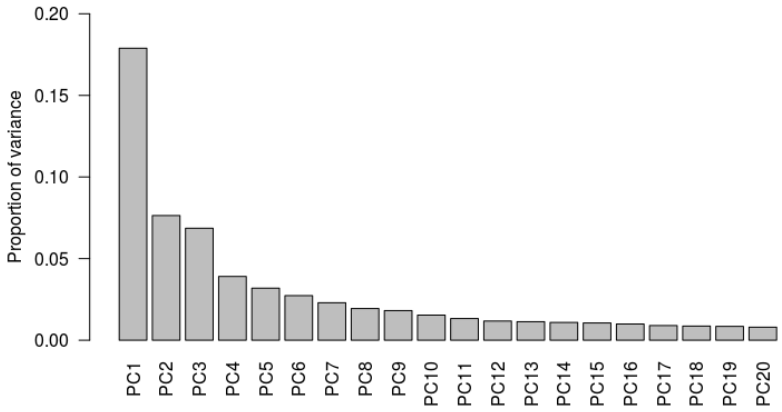


Figure AII.1: Fraction of variance explained by each principal component for a PCA where raw reads were grouped by sample. The first three PCs explain the biggest fraction of the variance and are therefore used as explanatory variables for the grouping plots.

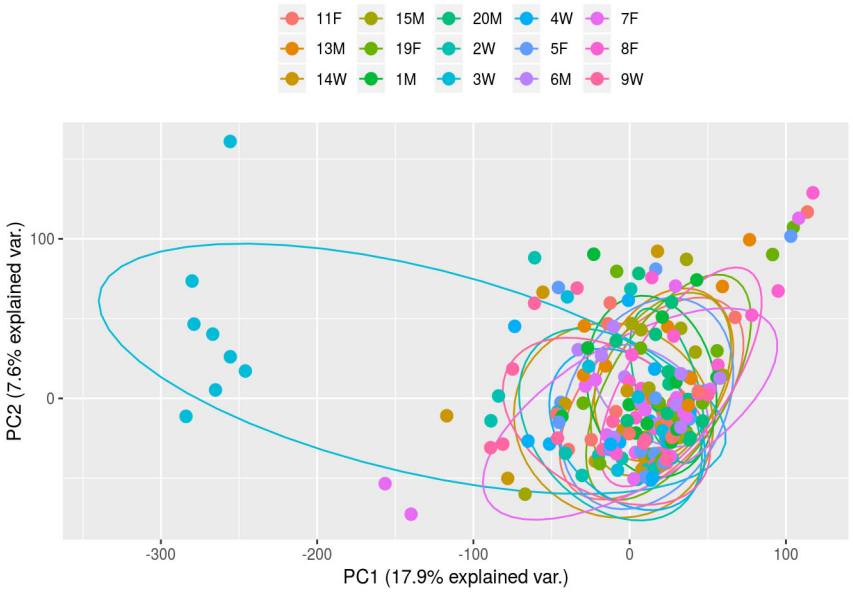


Figure AII.2: PCA analysis by sample, showing PC1 and PC2. Each dot in the plot is a different sample, *i.e.* a replicate of tissue and caste. In this case, samples are grouped by replicate, represented by the ellipses and the colour of the dots. A worker replicate (3W) seems to be problematic, as many samples from this replicate form a separate cluster.

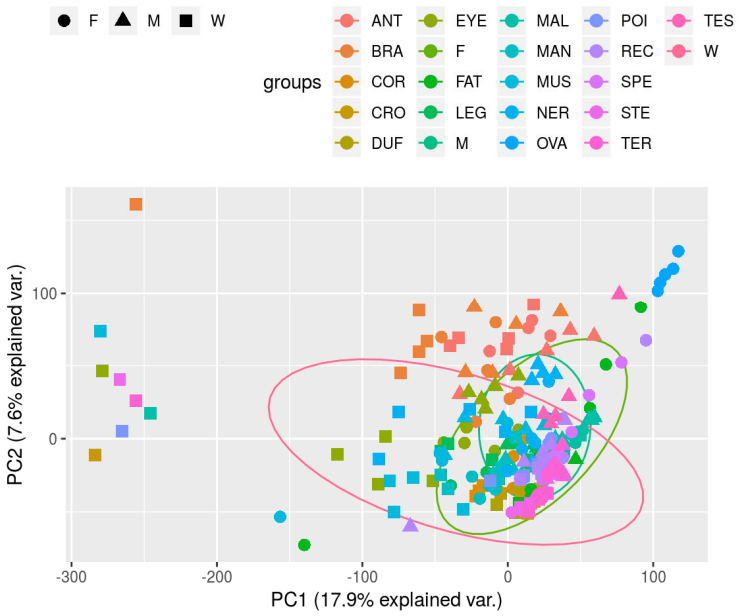


Figure AII.3: PCA analysis by sample, showing PC1 and PC2. Each dot in the plot is a different sample, *i.e.* a replicate of tissue and caste. In this case, samples are clustered by caste (M: males, circles, blue ellipse; F: female alates, triangles; green ellipse; W: workers, squares, pink ellipse) and tissue, represented by the colour of the dots. Males and females group together, whereas worker samples are more scattered.

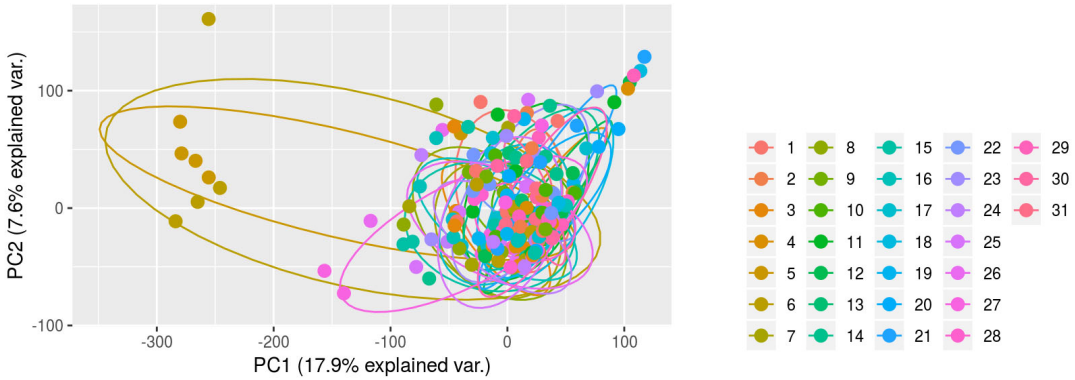


Figure AII.4: PCA analysis by sample, showing PC1 and PC2. Each dot in the plot is a different sample, *i.e.* a replicate of tissue and caste. In this case, samples are grouped by RNA extraction batch, which could affect the perceived expression patterns if, for example, a particular batch had a higher proportion of RNA degeneration. The PCA analysis does not show any clustering pattern by RNA extraction batch, indicating that this effect can be ignored in downstream analyses.

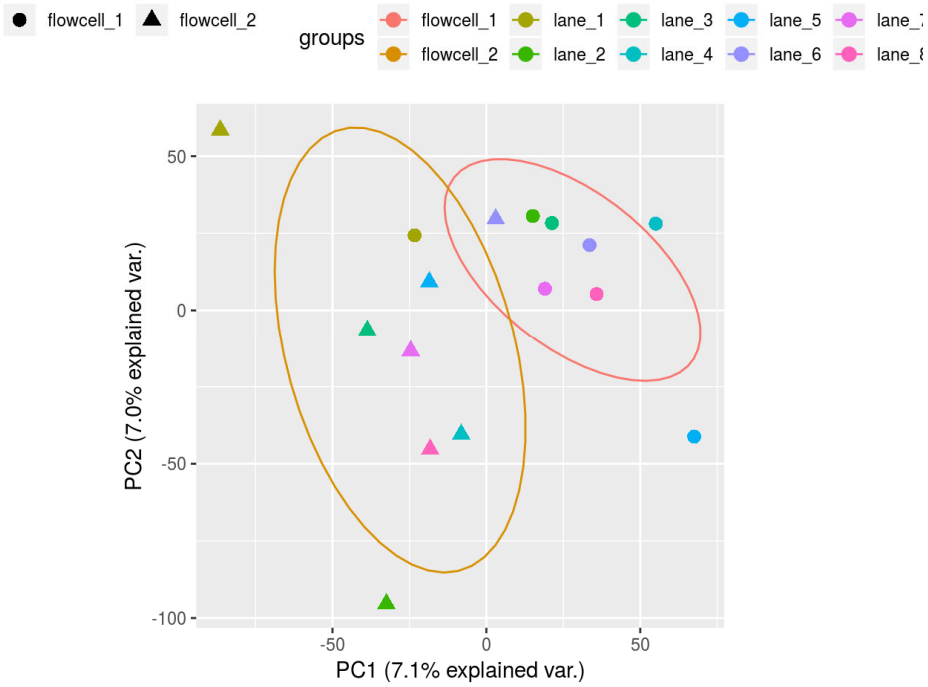


Figure AII.5: PCA analysis by sequencing lane, showing PC1 and PC2. Each dot in the plot is a different sequencing lane, *i.e.* one of the lanes where the libraries were sequenced in either flowcell 1 or 2. The plot shows a clustering by flowcell, indicating that, however small, there is an effect of sequencing flowcell in the perceived expression patterns that will need to be considered in

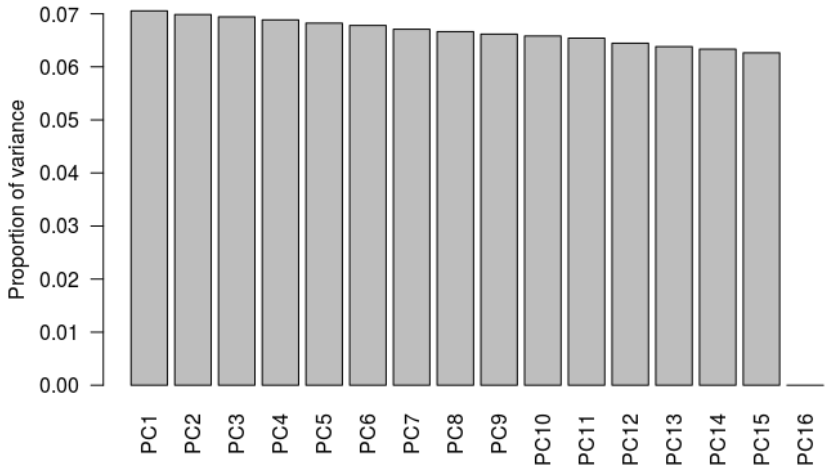


Figure AII.6: Fraction of variance explained by each principal component for a PCA where raw reads were grouped by flowcell lane. All PCs seem to explain a similar fraction of the variance, which implies that the underlying differences in variance caused by the explanatory variables (in this case, those arising from differences in sequencing) must be relatively small.

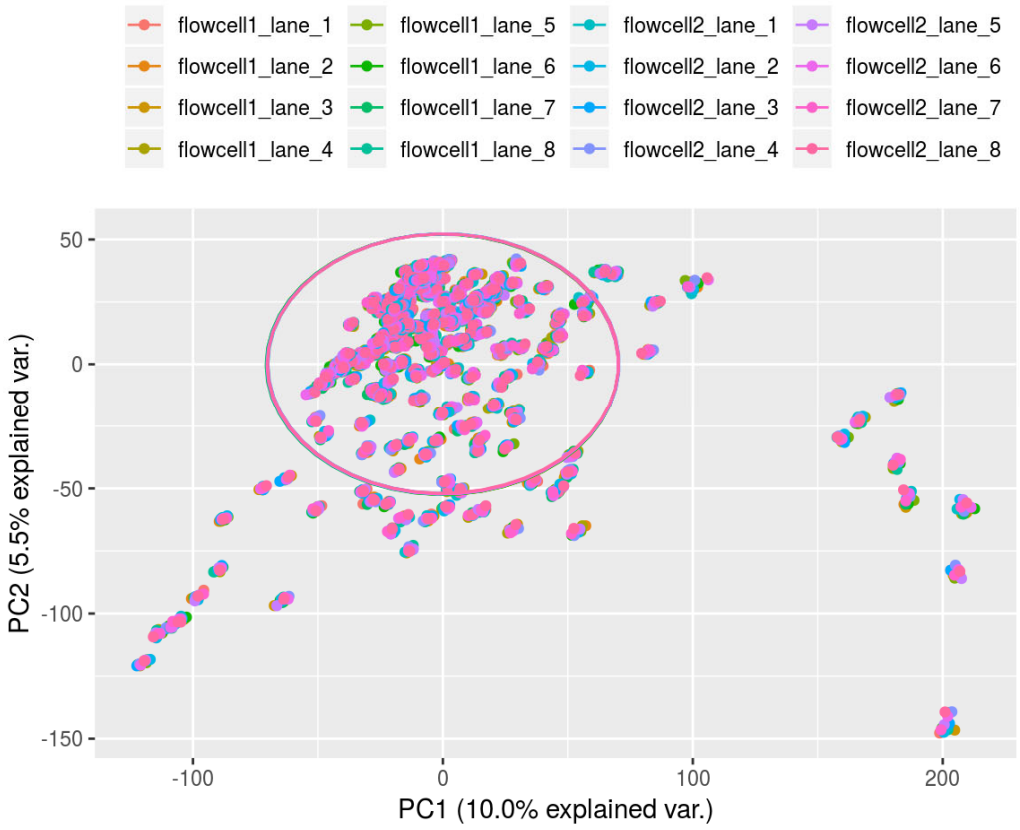


Figure AII.7: PCA analysis by individual raw reads file, showing PC1 and PC2. Each dot in the plot comes from a separate raw reads file, *i.e.* the reads generated per sample per lane within each of the two flowcells. This plot puts in perspective the relative effects of sequencing lane when compared to the effects arising from sample.

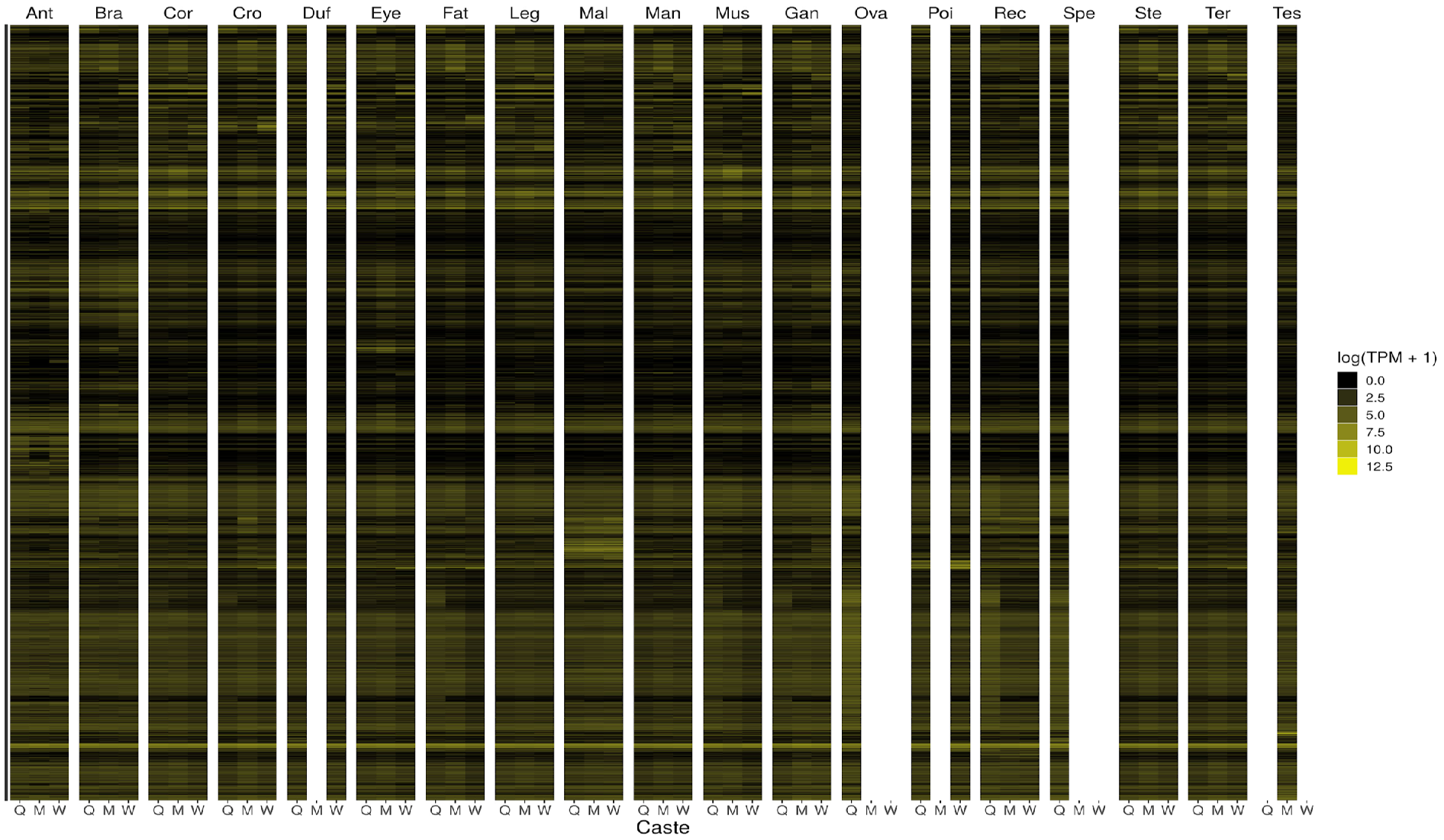

Ant
Bra
Cor
Cro
Duf
Eye
Fat
Leg
Mal
Man
Mus
Gan
Ova
Poi
Rec
Spe
Ste
Ter
Tes
Q M W
Caste
log(TPM + 1)
0.0
2.5
5.0
7.5
10.0
12.5

Figure AII.8: Heatmap showing the expression patterns of all genes available in the *S. invicta* reference across body parts (top axis) and castes (bottom axis). 84 genes were removed due to no expression in any of the samples. Each row represents one gene, each individual cell is a combination of gene, caste and tissue. Gene expression is measured as the logarithm of transcripts per million (TPM) + 1 for each gene. That is, the logarithm of the relative abundance of transcripts from each gene compared to the total of RNA molecules in the dataset. Note that some cells are empty, where the tissue was not available for a specific caste. The genes are ordered by similarity in expression patterns, that is, genes with similar expression patterns overall are plotted next to each other. Note that this visualisation allows to detect general expression patterns specific to tissues and caste. For instance, Malphigian tubules and midgut have a cluster of high gene expression in all castes that seems unique to this tissue. In other tissues with a strong differentiation per caste such as antennae or eyes, there are blocks of genes which show different expression patterns between castes. The three code letters for tissue correspond to: **Ant**, antennae; **Fat**, fat bodies; **Eye**, eye; **Man,** mandibles; **Leg**, legs; **Ter**, tergites; **Ste**, sternites; **Bra**, brain; **Mus**, muscle (thorax only); **Gan**, abdominal ganglia; **Cro**, crop; **Mal**, Malphigian tubules and midgut; **Rec**, rectum; **Duf**, Dufour gland (queens and workers only); **Poi**, Poison sac and gland (queens and workers only); **Spe**, spermatheca (queens only); **Tes**, testicles (males only); **Ova**, ovaries (queens only) and **Cor**, corpse. The caste labels correspond with **Q:** Queens, **M:** Males and **W:** Workers.

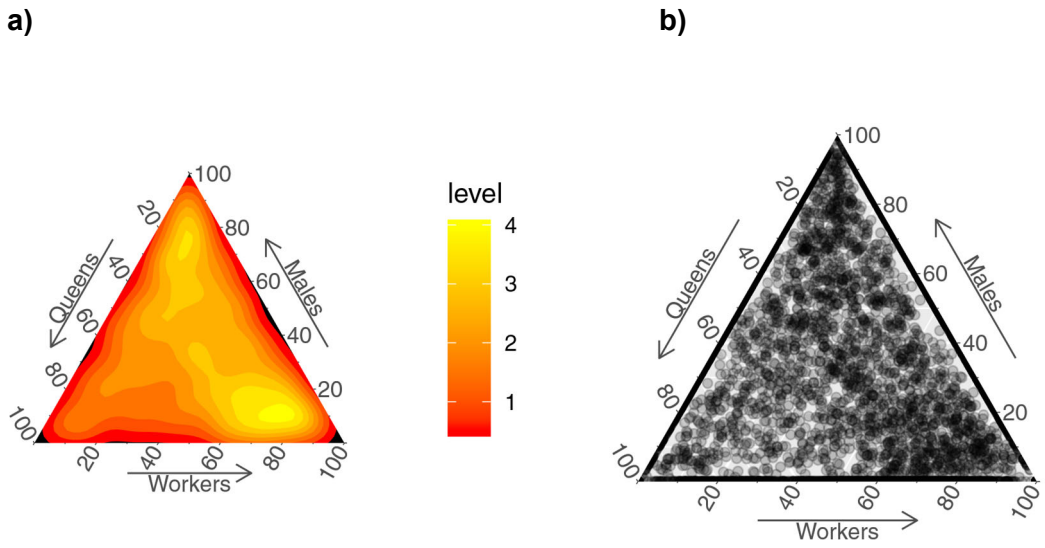


Figure AII.9. Caste-biased expression of genes. a) Each point represents a gene, with its location representing the relative expression in each caste, averaged across tissues. The value plotted along each axis (each side of the triangle) is 100% x log(e_{ij}/e_{ik}) /T_i, where e_{ij} is the relative expression of gene *i* in caste *j* and T_i is the total of the log ratio across all three comparisons. Hence, a value in the centre of the triangle represents a gene with equal expression levels in each caste, and a point displaced towards a vertex had relatively higher expression in the corresponding caste. The heatmap b) shows a higher density of genes plotted around the male and worker vertexes, illustrating a greater number of male and worker biased genes (relative to queen biased genes). Very few genes had expression essentially limited to one caste (there is a low density of points around the extreme vertices).

Annex III: Related publications

Attached are list of publications I was involved in during my PhD:

Martínez-Ruiz, C., & Knell, R. J. (2017). Sexual selection can both increase and decrease extinction probability: reconciling demographic and evolutionary factors. *Journal of Animal Ecology*, *86*(1), 117-127.

Knell, R. J., & Martínez-Ruiz, C. (2017). Selective harvest focused on sexual signal traits can lead to extinction under directional environmental change. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1868), 20171788.

Pracana, R., Levantis, I., Martínez-Ruiz, C., Stolle, E., Priyam, A., & Wurm, Y. (2017). Fire ant social chromosomes: Differences in number, sequence and expression of odorant binding proteins. *Evolution letters*, *1*(4), 199-210.

Favreau, E., Martínez-Ruiz, C., Santiago, L. R., Hammond, R. L., & Wurm, Y. (2018). Genes and genomic processes underpinning the social lives of ants. Current opinion in insect science, 25, 83-90.