

DR JACK HOWE (Orcid ID : 0000-0001-9126-4471)

Article type : Research Papers

Corresponding Author Email ID: jack.howe@zoo.ox.ac.uk

**A novel method for using RNA-seq data to identify imprinted genes in social Hymenoptera with multiply mated queens**

**Abstract**

Genomic imprinting results in parent-of-origin dependent gene expression biased towards either the maternally- or paternally-derived allele at the imprinted locus. The kinship theory of genomic imprinting argues that this unusual expression pattern is a manifestation of intra-genomic conflict between the maternally- and paternally-derived halves of the genome that arises because they are not equally related to the genomes of social partners. The theory thus predicts that imprinting may evolve wherever there are close interactions among asymmetrically related kin. The social Hymenoptera with permanent caste differentiation are suitable candidates for testing the kinship theory because haplodiploid sex determination creates strong relatedness asymmetries and nursing workers interact closely with kin. However, progress in the search for imprinted genes in the social Hymenoptera has been slow, in part because tests for imprinting rely on reciprocal crosses that are impossible in most species. Here, we develop a method to systematically search for imprinting in haplodiploid social insects without crosses, using instead samples of pooled individuals collected from natural colonies. We tested this protocol using data available for the leaf-cutting ant *Acromyrmex echinator*, providing the first genome-wide search for imprinting in any ant. While we identified several genes as potentially imprinted, none of the four genes tested

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi:](#)

[10.1111/jeb.13716](#)

This article is protected by copyright. All rights reserved

could be verified as imprinted using digital droplet PCR, highlighting the need for higher quality genomic assemblies that accurately map duplicated genes.

## Introduction

Genomic imprinting describes the allele-specific expression (ASE) of a gene dependent on the parent it was inherited from, based on epigenetically inherited factors such as DNA modifications or potentially small RNAs (Reik and Walter, 2001). Often, genomic imprinting is observed as complete silencing of one allele and expression exclusively from the other, but imprinted genes that show only an incomplete bias are increasingly being found (eg. Galbraith et al., 2016, Smith et al., 2020). While there are a number of theories that propose explanations for the evolution of imprinting (reviewed in Haig, 2014, Spencer and Clark, 2014), the best supported of these—the kinship theory of genomic imprinting—posits that genomic imprinting arises because of conflict between maternally- and paternally-inherited alleles. The kinship theory is particularly well supported for genes expressed in tissues that affect resource provisioning to offspring in mammals (Haig, 2004) and, because asymmetrical relatedness underpins imprinting in these systems, the kinship theory has also predicted genomic imprinting to occur in haplodiploid hymenopteran social insects (Queller, 2003, Kronauer, 2008). Yet, while several empirical approaches have accumulated partial genetic and phenotypic support for this contention (reviewed in Pegoraro et al., 2017), progress has been hampered by difficulties in rearing and crossing social insect models other than honeybees and bumblebees (eg. Kocher et al., 2015, Galbraith et al., 2016). The work reported in this paper exposes these empirical challenges, and proposes an alternative approach that requires high quality reference genomes, but is independent of most aspects of experimental rearing.

The kinship theory of genomic imprinting is an extension of inclusive fitness theory. In its most general form, inclusive fitness theory states that any gene within an individual is selected according to both its effects on the individual's direct reproductive success and its effects on the reproductive success of its social interactants, weighted by the probability that they carry the same gene, i.e. their relatedness (Hamilton, 1964a, 1964b, West et al., 2007). Thus, a gene with direct fitness cost will still spread if it has a sufficiently positive fitness effect on the reproductive success of related individuals. Relatedness is often discussed in 'whole-individual' terms, because honest meiosis generally keeps the interests of different genes aligned (Grafen, 1985). However, as highlighted by the kinship theory of genomic imprinting, maternally-derived genes (matrigenes) and paternally-derived genes (patrigenes) are often not equally likely to be found in kin, and will

thus value kin differently and be selected in different directions depending on the role they find themselves in (Haig, 2000).

The kinship theory is most often discussed in relation to cross-placental resource allocation conflict in mammals where, in mating systems that are not strictly monogamous, patrigenes are less likely than matrigenes to contribute to a mother's future offspring. The patrigenes, therefore, will be more 'selfish', and will favour diverting more resources from the mother to a focal foetus than the matrigenes do (Moore and Haig, 1991). Consistent with this notion, patrigenes expressed in the placenta generally promote, while expressed matrigenes compensate for this and inhibit, foetal growth (Haig, 1996, 2004). While the kinship theory is well-supported in mammals, the logic applies equally wherever there are relatedness asymmetries and close interactions among kin, such as in the endosperm of vascular plant seeds (Haig and Westoby, 1989, 1991, Wolff et al., 2011, Gehring and Satyaki, 2017), when there are sex- or age-specific dispersal asymmetries (Úbeda and Gardner, 2010, 2011, 2012) or, pertinent to our present study, in haplodiploid social insects where provisioning investments are made by 'somatic' life-time un-mated siblings that are functionally comparable to the somatic tissues that mediate provisioning in mammals and plants (Queller and Strassmann, 2002, Queller, 2003, but see Kronauer, 2008). Imprinting is thus predicted to possibly occur in the superorganismal ants, corbicular bees and vespine wasps that have obligately differentiated nursing castes, while predictions are less clear in cooperatively breeding social insects where helper altruism remains facultative (Boomsma and Gawne, 2018) and an individual's relatedness to nestmates is more dynamic over time (but see Queller and Strassmann, 2002, Queller, 2003).

The colonial life-histories of the advanced social insects imply that potentially millions of related individuals interact closely throughout their lives, while their haplodiploid sex determination system creates strong relatedness asymmetries within a colony (Hamilton, 1964a, Trivers and Hare, 1976, Boomsma and Grafen, 1991). If a single queen founds a colony after mating with multiple males, as in the honeybees (Withrow and Tarpay, 2018) or the leaf-cutting ants (Villesen et al., 2002), the diploid female workers within the colony will, on average, be more closely related through their matrigenes than through their patrigenes. This asymmetry is predicted to produce imprinting conflict for many aspects of colony life (Queller, 2003), such as caste determination (Dobata and Tsuji, 2012), sex ratio (Wild and West, 2009) and worker reproduction,



depending on queen number and queen-mating-frequency (Queller, 2003). These social Hymenoptera therefore appear to offer a robust, independent testbed for the kinship theory of genomic imprinting (Queller and Strassmann, 2002), even though this expectation rests on a series of implicit assumptions that may not be fulfilled (Kronauer, 2008).

With a solid theoretical basis for expecting genomic imprinting in the superorganismal social Hymenoptera, the search for parent-of-origin effects consistent with the kinship theory has indeed been accumulating suggestive evidence (reviewed in Pegoraro et al., 2017). Consistent with the kinship theory, disruption of DNA-methylation associated with imprinting promotes worker reproduction in a bumblebee, leading to increased aggression and ovary size (Amarasinghe et al., 2014), while the genes involved appear mono-allelically methylated and expressed (Amarasinghe et al., 2015, Lonsdale et al., 2017). In an ant, parent-of-origin effects influence development into queens rather than workers (Libbrecht et al., 2011) and brood care behaviour in workers (Libbrecht and Keller, 2013). Finally, reciprocal crosses between honeybee subspecies have demonstrated parent-of-origin effects on aggression (Guzman-Novoa et al., 2005) and ovary size (Oldroyd et al., 2014). However, as highlighted elsewhere, asymmetrical expression of these traits is also consistent with cyto-nuclear-incompatibilities (Gibson et al., 2015).

Only five social insect studies have attempted to directly identify putatively imprinting-induced ASE with genome-wide searches based on high-throughput sequencing: three in the honeybee (Kocher et al., 2015, Galbraith et al., 2016, Smith et al., 2020), and two in a bumblebee (Lonsdale et al., 2017, Marshall et al., 2020). These studies adopted techniques successfully applied in other (diploid) organisms, where two genotypes are reciprocally crossed and offspring RNA sequenced to find genes where the relative proportion of the two RNA alleles is dependent on the direction of the cross while differing significantly from the 1:1 null expectation (Wang et al., 2008, Gregg et al., 2010, Gehring et al., 2011, Coolon et al., 2012, Babak et al., 2015). However, the application of this approach to the social insects is not straightforward for two reasons. First, a hymenopteran colony contains a minimum of three (but often many more) haploid genomes, complicating tests for parent-of-origin effects. Studies therefore created unnaturally simple colonies by singly inseminating queens—honeybee queens are naturally inseminated by 10-20+ drones (Withrow and Tarpy, 2018)—to create an approximately diploid pool for loci where the queen is homozygous (Kocher et al., 2015, Galbraith et al., 2016, Smith et al., 2020). Second, social insect colonies

generally have prohibitively long generation times—a problem only partially alleviated in annual bumblebees, and with artificial insemination and queen rearing in the honeybees—so controlled breeding is either laborious or, more often, impossible.

Here, we develop an alternative approach that extends the search for genomic imprinting beyond the bees. We outline a graphical model to identify potentially imprinted genes without using reciprocal crosses or artificial insemination, but pooling diploid offspring and assessing colony-level ASE patterns. Although this method has lower statistical power than the reciprocal cross technique, it can be applied to any species with good quality genomic resources, so long as workers can be collected and the maternal genotype can be determined, either by direct assessment or from assessing her haploid sons. We apply this model to a dataset produced previously for the leaf-cutting ant *Acromyrmex echinator* (Li et al., 2014), which allowed us to conduct the first genome-wide search for imprinted genes in any ant. Finally, we use allele-specific qPCR to verify parent-of-origin ASE for four genes that our screening highlighted as potentially imprinted, to find that none of them could be proven to show ASE consistent with imprinting.

## Methods

### The imprinting model

Our model is based on the same principles used in diploid organisms. Under typical unimprinted expression, we expect parity between the allelic proportions in the DNA and the RNA, but genomic imprinting should result in significant deviations from parity (Wang and Clark, 2014).

The model differs however, in that it considers the effects of imprinting on pools of diploid individuals collected from a hymenopteran colony comprising at least three haploid genomes (a diploid queen and  $\geq 1$  haploid male) rather than just a simple diploid genome. This model was developed to analyse data of the type published by Li et al. (2014), where RNA and DNA extracted from the same pools of individuals were sequenced to investigate and confirm that caste-specific RNA-editing occurred in the leaf-cutting ant *Acromyrmex echinator*. The data that we used are described in more detail below where we discuss the application of the model.

In whole-genome screens for genomic imprinting, alternative bases at a single nucleotide polymorphism (SNP) in an RNA- and DNA-sequence identify the two parental alleles. Any SNP in a diploid organism thus represents a heterozygous locus. This implies that a focal and

alternative ‘SNP-allele’ should be present in equal proportions in the DNA and, if the gene is not imprinted, also in the RNA. However, things are more complex in a pool of diploid individuals collected from a social insect colony: here, there is the diploid genome inherited from the mother queen, and a series of  $n$  haploid genomes, where  $n$  is the number of males that inseminated her. In the most simple colonies, only a single male has contributed but, in more complex social systems like the honeybees, many males may be represented (eg. Withrow and Tarpy, 2018).

Consequently, the mother and any of the fathers may carry an alternative SNP-allele so that, in a pool of diploid offspring collected from a colony, the frequency of a given SNP-allele can take any value between 0 and 1, rather than being restricted to 0.5 as in a single diploid organism. The frequency of a SNP-allele in DNA extracted from such a pool is the average of the SNP-allele frequency in the queen ( $q$ ) and the weighted SNP-allele frequency ( $m$ ) contributed by the  $n$  males she mated with, with the queen’s frequency fixed at 0.5 if she is heterozygous at a locus, or at 0 or 1 if she is homozygous:

$$DNA = \frac{q}{2} + \frac{1}{2n} \sum_{i=1}^n m_i \quad (1)$$

Parent-of-origin dependent ASE will lead to the frequency of a SNP-allele in the colony RNA to deviate from its frequency in the colony DNA, and will cause the RNA SNP-allele frequency to more closely resemble either the maternal or paternal DNA. According to the ‘loudest-voice-prevails principle’, imprinting conflict should ultimately lead to mono-allelic expression of the affected genes (Haig, 1996, Burt and Trivers, 2006). Under such mono-allelic expression, the frequency of a SNP-allele in the colony RNA will then reflect either the genotype of the queen or the weighted average of the genotypes of the males she mated with. Many imprinted genes in mammals (Haig, 2004), and a smaller number of genes in bumblebees (Lonsdale et al., 2017), show monoallelic expression, but it is unclear whether the ‘loudest-voice’ principle is universally applicable as incomplete ASE biases are often found, particularly in the social insects (eg. Kocher et al., 2015, Galbraith et al., 2016, Smith et al., 2020). Where silencing is incomplete, the colony RNA will be intermediate between the parental genotypes.

### ***Predictions under matrigenic, patrigenic and incomplete imprinting***

Under exclusively matrigenic expression, the colony-level offspring (workers and gynes) RNA will reflect the maternal DNA. Because the queen is diploid, the frequency of a SNP-allele at any locus in her genome is 0 or 1 when she is homozygous, or 0.5 when she is heterozygous. As the queen contributes half of the DNA of her female offspring, her genotype sets upper and lower bounds to the possible frequency of a SNP-allele among her daughters in the colony. Thus, under matrigenic expression, there are three expectations for the relationship between the DNA and RNA in a pool of diploid female offspring corresponding to her genotype. First, if the queen is homozygous for a focal SNP-allele, the frequency of that SNP-allele in the pooled DNA of her daughters will lie somewhere between a minimum of 0.5 if all the males she mated with carried the alternative SNP-allele and a maximum of 1 if all males carried the same SNP-allele as she does, while the frequency of the SNP-allele in the colony RNA must be 1. Second, if the queen is homozygous for the alternative SNP-allele, the expectations are reversed: the SNP-allele frequency in her daughters' DNA pool must then be less than 0.5, and the SNP-allele should be absent in the RNA. Finally, if she is heterozygous, the DNA SNP-allele frequency in the pool of daughters will lie between 0.25 and 0.75, and the RNA SNP-allele frequency will be 0.5. These predictions are illustrated in Figure 1a, and are described mathematically as:

$$RNA_m = q \quad (2)$$

to quantify that the colony-level RNA SNP-allele frequency under matrigenic expression ( $RNA_m$ ) reflects the queen's genotype ( $q$ ).

Similarly, there are three complementary predictions under exclusively patrigenic expression. In this scenario, the colony-level (workers and gynes) RNA proportions will reflect the paternal DNA, that is, the RNA SNP-allele frequencies should equal those of the males that contributed to the queen's offspring, weighted by their relative contributions to the colony's offspring. The queen's genotype sets the same limits on the possible DNA proportions as described above, but unlike the matrigenic predictions, colony-level SNP-allele frequency for RNA can now take any value between 0 and 1. These predictions are illustrated in Figure 1b, and can be described mathematically as:

$$RNA_p = m = 2 \cdot DNA - q \quad (3)$$

where the colony RNA under patrigenic expression ( $RNA_p$ ) reflects the weighted average genotypes of the contributing males ( $m$ ). Also here, the colony allele-frequency for  $DNA$  is calculated as in equation (1), but now the slope of the DNA-RNA relationship equals 2 because, under patrigenic expression, the fathers contribute all of the RNA at an imprinted locus but only half of the DNA.

Although the loudest-voice prevails principle may hold for many imprinted genes, it is clear that it does not hold for many others. In those cases, an average individual's (and therefore the colony's) RNA will reflect a mix of the maternal and paternal genotypes, depending on the degree of bias. To accommodate for this usually unknown variation, the above models can be generalised by summing the matrigenic and patrigenic imprinting expectations, and weighting them by a 'relative degree of bias' factor towards matrigenic imprinting, which we denote as  $\alpha$ , so we obtain the expression:

$$RNA = \alpha q + (1 - \alpha)(2DNA - q) \quad (4)$$

Here, an unimprinted locus would have no bias towards the matrigenic ( $\alpha = 0.5$ ) so the RNA proportion remains equal to the DNA proportion, while deviations towards  $\alpha = 0$  and  $\alpha = 1$  will make equation (4) converge towards equations (3) and (2), respectively. Often, an arbitrary cut-off of 0.65 (0.35) is used to define when a gene is considered imprinted (Wang and Clark, 2014), but we use a more stringent cut-off of 0.8 (0.2) here.

### **Testing the model using DNA and RNA data from the leaf-cutting ant *Acromyrmex echinator***

Testing the model outlined in the previous section requires pairwise colony-level estimates of SNP frequencies for both DNA and RNA extracted from the same pools of female offspring. These data were available from a previous study (Li et al., 2014). To briefly summarize, Li et al. (2014) sampled pools of 200 small workers, 50 large workers, and 50 winged reproductive females (gynes) from three colonies (Ae322, Ae356 and Ae363). RNA was extracted from the pooled

heads of individuals of each caste, separately for each colony, and DNA was extracted from the remaining bodies, restricting the study to gene expression in the brain but providing the colony-level dataset required to test our present model. DNA and RNA were both sequenced using the Illumina HiSeq platform, and reads were aligned to the *A. echinator* genome (Nygaard et al., 2011) using the Burrows-Wheeler Aligner (Li and Durbin, 2009). Each end of the uniquely mapping reads was trimmed by 6bp and then realigned to the genome using the BLAST-like alignment tool (BLAT) (Kent, 2002), and only those reads that were supported by both alignment methods were retained for further analysis.

For the purpose of the present study, we developed a pipeline to identify sites with significant allele-specific expression. First, any low-quality reads (quality score  $< 20$ ) were discarded, as were loci to which fewer than 10 reads mapped. The Bayesian statistical framework in RES-Scanner (Wang et al., 2016) was used to identify all polymorphic loci (SNPs). This computes the probability of each possible genotype using the observed data (i.e. the mapped bases and corresponding sequencing qualities) for each position on the reference genome, after which the genotype with the highest probability is reported (see additional file 1 of Wang et al., 2016). Only those sites with a posterior probability  $> 0.95$  were retained as confirmed SNPs. Fisher's exact tests were then used to evaluate differences in the RNA and DNA allele frequencies for all SNPs, and SNPs were identified as showing significant allele-specific expression (ASE) when they met a significance level of  $p < 0.05$  after correction for false discovery rate.

For each SNP where at least one sample showed ASE, the data across all nine samples (three castes across three colonies) were tested for consistency with the imprinting expectations obtained from our model (Figure 1). Confidence intervals (Agresti and Coull, 1998) were constructed for the allele frequency in the DNA based on the number of reads supporting each of the two alternative SNP alleles. The minimum and maximum (95%) bounds of these confidence intervals were then used to predict RNA ranges under all six maternal and paternal imprinting scenarios, using equations 1-4. The same kind of confidence intervals were generated for the observed proportions in the RNA-sequencing reads, allowing us to test for overlap between these observed RNA confidence intervals and the expected RNA intervals predicted from the DNA-sequencing data under the various imprinting scenarios. Finally, we used a combination of criteria to establish whether expression at a locus was consistent with imprinting predictions: we required that all

samples for each locus were consistent with the same direction of imprinting (either maternal or paternal), and that the predicted queen genotype for each SNP was the same within a colony (but may differ between colonies). We thus obtained a subset of all identified SNPs where at least one of the nine samples was inconsistent with the unimprinted null expectations (i.e. with significant ASE) and where all nine samples were consistent with predicted allelic ratios under biased expression of one of the parental alleles ( $\alpha \geq 0.8$  and  $\alpha \leq 0.2$ ). That is, all points fit the predicted lines in Figure 1, although they did not all necessarily show significant ASE as the imprinted predictions for some DNA SNP-allele frequencies overlapped with unimprinted predictions (where lines cross in Figure 1).

### **Validation of allele-specific expression data using ddPCR**

#### *Sample collection and RNA and DNA extractions*

Six *A. echinator* colonies were used in this follow-up study (Ae160B, Ae168, Ae226, Ae263, Ae322, Ae356), two of which (Ae322 and Ae356) were also sampled for the construction of the original dataset (Li et al. 2014). Colonies were collected in Gamboa, Panama between 2001 and 2008 and had since been maintained in Copenhagen at 25°C and ca. 70% relative humidity, on a diet of bramble leaves, apple and rice. Four samples were collected from each colony: 50 gynes, 50 large workers, 100 small workers and 50 males. At the time of collection, some colonies were not producing both male and female reproductives in sufficiently high numbers, so these data points were omitted, meaning that pooled gynes were only available for colonies Ae168, Ae263 and Ae356, while no males were available for colony Ae226. Ants were flash frozen using liquid nitrogen during collection, after which the head and body of each ant were separated using forceps, and pooled separately according to caste and colony. Both pools were then ground to a fine powder in liquid nitrogen using a pestle and mortar. We can be reasonably certain that the males collected were queen- rather than worker-derived, because worker reproduction is suppressed in *A. echinator* while the queen is present (Dijkstra et al., 2005). To validate this assumption, we extracted DNA from individual males of three colonies (Ae226, Ae322, and Ae356) whose workers had previously been genotyped using eight microsatellite loci so maternal and paternal genotypes were known, and found the male-offspring genotypes to be consistent with being queen-produced.

Of the homogenised pooled heads, a sample of approximately 10mm<sup>3</sup> was used for RNA extractions using the QIAGEN Universal Mini Kit following the manufacturer's protocol. For the small workers this was the entirety of the sample, and for the large workers and gynes this represented about one third of the available biomass. We also included an on-column DNase digestion step during RNA extractions to prevent carry-over of genomic DNA that could influence allelic ratios. DNA was extracted from approximately 20mm<sup>3</sup> of the matching homogenised body pools using the QIAGEN DNeasy kit, following the manufacturer's protocol. The success of DNA and RNA extractions was checked using a NanoDrop spectrophotometer.

RNA was reverse transcribed to cDNA in 10µl reactions containing 5.25µl of the RNA sample, 0.5µl SuperScript III (Invitrogen) 2µl 5X first strand buffer, 1µl DNTP, 1µl DTT, 0.125µl RNASin, and 0.125µl of the random primer Qt. RNA was heated to 65°C for 3 minutes before being placed on ice and mixed with the reagents. This was then heated to 42°C for 60 minutes, 50°C for 10 minutes, and finally 70°C for 15 minutes. Following reverse transcription, the cDNA samples were treated with 1.5 units RNase H and incubated at 20°C for 20 minutes. These samples were then diluted to approximately 3.5ng/µl based on RNA quantity assessment with a NanoDrop spectrophotometer. As recommended for ddPCR applications, we digested 8.5 µL of the DNA extractions using 0.5 µL of the *HindIII* restriction endonuclease with 1 µL digestion buffer at 30°C for 20 minutes, followed by 10 minutes at 80°C, after which digested products were diluted in 190µL water. None of the regions targeted during PCR reactions contained a *HindIII* restriction site.

### ***Allele-specific quantitative-PCR***

The relative frequencies of the two alleles at the SNPs of interest were assessed using competitive allele-specific (CAST) digital droplet PCR (ddPCR) assays—which rely on two fluorescently-marked Taqman probes, each of which preferentially binds to one SNP allele. Primers and fluorescent probes were designed for four genes using the web-based Primer3Plus (Untergasser et al., 2012) to span an approximately 100bp region including the locus-specific SNPs of interest. Two probes were designed for each locus, one for each alternative SNP, of which one was labelled with HEX and the other with FAM. The exact position of each probe was designed to maximise the difference in temperature between annealing to the correct versus the alternative allele using



the web-based OligoAnalyzer from Integrated DNA Technologies (<https://eu.idtdna.com/calc/analyzer>). Primers and probes were designed for four genes of the 46 that fitted ASE expectations under genomic imprinting: *Major Royal Jelly Protein 3*, *Vitellogenin 1*, *Histone Acetyl-Transferase SETMAR*, and *S1 RNA-binding protein*. The primer and probe sequences are given in table S1. In more detail, these genes were chosen according to their fit to imprinting predictions—based on minimising the total absolute perpendicular residuals between imprinting predictions and the sequencing data for each locus—and the feasibility to design primer and probe sets for a locus. This required that the focal gene was not obviously duplicated, and that primers and probes did not form strong dimer pairs. Other necessary conditions were that primers were specific to the region of interest, targeted an individual SNP, and had no additional SNPs falling within the primer-binding-regions on either side. This is because any SNPs within the binding region of the primer could have systematically biased ddPCR results towards one of the two alleles when they were within circa max 50bp of the focal SNP and therefore likely linked. Such linkage would affect the relative binding temperatures and allele-specific reaction efficiency.

Digital droplet PCR reactions consisted of 10µl ddPCR SuperMix for probes (no dUTP), 1.8 µl of each forward and reverse primer (10µM solutions), 0.5 µl of the two Taqman probes (10µM stock solutions), 5.4 µL water and 2 µL DNA or cDNA template. This reaction solution was then split into typically 15-20,000 droplets using the BioRad droplet generator, before PCR reactions were conducted using a two-step thermal protocol: an initial denaturation of 95°C for 10 minutes, followed by 40 cycles of 94°C for 30s and a 90 second annealing step at a gene-specific temperature (Table S1), with a final signal stabilization step at 98°C for 10 minutes. Plates were analysed immediately after the PCR reactions using the BioRad Droplet Analyzer.

We conducted additional ddPCR reactions with the same primers to assess the number of copies of each gene. These reactions differed from those described above in that they contained ddPCR SuperMix for EVA-Green rather than the SuperMix for Probes, and did not contain the fluorescent probes. Reactions were conducted on a mixture of all extracted DNA samples, which was serially diluted seven times by a factor of two. An additional reference gene was included as a standard against which the genes of interest could be compared (*TATA-box protein*, sequence shown in table S1), which our tests indicated was present as only a single copy. These reactions consisted of 10 µl ddPCR SuperMix for EVA-Green, 0.5 µL of the forward and reverse primers, 9 µL water

and 2  $\mu$ L DNA template. Droplets were generated and reactions conducted as described above with a 58°C annealing temperature.

## Results

Our model proposed seven possible expression patterns for any locus within a colony: a null-hypothesis of unimprinted expression where RNA and DNA frequencies do not differ (dashed diagonals in Figure 1), and six alternative imprinting scenarios (horizontal lines in Figure 1a, and steep diagonals in Figure 1b, and the shaded areas around each). The final testing dataset consisted of 517 SNPs which showed ASE in at least one sample. We restricted further evaluation to those that were located within genes with putative functions, which left 168 SNPs distributed across 120 genes. Of these, 92 SNPs across 43 genes showed DNA and RNA SNP-allele frequencies consistent with genomic imprinting at an  $\geq 80\%$  bias towards one parental allele: 55 SNPs across 24 genes with a matrigenic bias (table S2), and 59 SNPs across 36 genes with a patrigenic bias (table S3). 22 SNPs were consistent with both matrigenic and patrigenic expression, as predictions overlap in some of the parameter space (Figure 1). When restricted to complete silencing of one allele and exclusive expression from the other, 63 SNPs across 34 genes were consistent with expectations under imprinting: 25 SNPs across 10 genes with matrigenic expression ( $\alpha = 1$ ) and 47 SNPs across 31 genes with patrigenic expression ( $\alpha = 0$ ). Relaxing the assumption that imprinting should be consistent across all castes showed—as expected—a greater number of SNPs consistent with imprinting in the two worker castes with the gynes omitted. This worker-specific analysis identified 72 SNPs across 45 genes consistent with an incomplete patrigenic bias ( $\alpha \leq 0.2$ ) and 63 SNPs in 44 genes consistent with complete bias ( $\alpha = 0$ ), as well as 67 SNPs across 27 genes consistent with an incomplete matrigenic bias ( $\alpha \geq 0.8$ ), and 30 SNPs across 11 genes consistent with a complete bias ( $\alpha = 1$ ). The candidate loci were located within genes with a range of functions, including transcription factors, histone-modifying enzymes, and genes associated with phenotypic reproductive division of labour. These lists, however, contained no significantly enriched functional groups or pathways.

Consistency in the putative imprinting status across loci within a gene can provide further support for the imprinting status of that gene. However, 43 of the 120 genes tested here contained only a single informative SNP (Table S4), as loci that do not show significant colony-level ASE cannot generally be distinguished from unimprinted expression and are therefore not informative under

this method. For genes with loci identified as consistent with a complete patrigenic bias, 15 had a single SNP, while only three of the remaining 17 with multiple SNPs were internally consistent (each containing only two informative SNPs). Likewise, with complete matrigenic bias, 3 of 11 genes contained only one informative SNP, and at all others, loci within a gene were found to disagree (Table S4). Often, only a single of the nine tested samples was inconsistent at a given locus (Table S5, Figure S1), but the exact sample that was inconsistent with imprinting varied among the loci within a gene (data not shown). Samples where we identified significant colony-level ASE at a given SNP were more likely to be the ones that caused inconsistency with imprinting scenarios in all cases except under exclusive matrigenic expression (ie. an  $\alpha$ -value of 1) (Fisher's exact tests,  $p < 0.01$ ). While inconsistencies among SNPs within loci suggest that these genes are likely not imprinted, we did not consider this necessarily fatal to the possibility of imprinting at such loci due to the possibility of alternative transcripts or other issues with genome assembly that were beyond the scope of this study.

We used the total residuals between imprinted expectations and SNP-allele frequencies to select four loci for further analysis: two genes consistent with exclusive matrigenic expression (*Histone-lysine N-methyltransferase SETMAR* and *Vitellogenin 1*), and two consistent with exclusive patrigenic expression (*Major Royal Jelly Protein 3* and an *S1 RNA-binding domain containing protein*). *Histone-lysine N-methyltransferase SETMAR* and *Vitellogenin 1* each contained multiple informative SNPs of which several, although not all, were consistent with imprinting. *Vitellogenin 1* contained five SNPs, two were consistent with both matrigenic and patrigenic expression, one had only a single inconsistent sample, while the other SNPs separated by  $> 6000$  bases had more inconsistent samples (Table S5). *Histone-lysine N-methyltransferase SETMAR* contained eight SNPs split across two scaffolds: of the two on scaffold 361 where the SNP of interest was located, only one of the two was consistent with imprinting (Table S5). *Major Royal Jelly Protein 3* and an *S1 RNA-binding domain containing protein* contained only a single informative SNP (Table S4, S5).

The DNA and RNA allelic-proportions estimated by ddPCR were consistent with those obtained from the sequencing data for three genes (*Major Royal Jelly Protein 3*, *S1 RNA-binding domain containing protein*, and *Histone-lysine N-methyltransferase SETMAR*; Figure 2), but for *Vitellogenin 1*, ddPCR-estimated allele frequencies differed substantially from the sequencing data

(Figures 2A *versus* 2B). This discrepancy suggested that the *Vitellogenin 1* reactions were unreliable, and, as we were unable to design alternative probes, this gene was not included in further investigations. While all four genes showed expression consistent with imprinting as estimated by direct sequencing, the ddPCR results indicated that *SI RNA-binding protein* had DNA-RNA relationships that best fitted non-imprinted predictions (Figure 2B), while the two remaining genes, *SETMAR* and *Major Royal Jelly Protein 3*, continued to match imprinted expectations. Allele frequencies for *Major Royal Jelly Protein 3* were consistent with patrigenic expression, and those of *Histone-lysine N-methyltransferase SETMAR* were consistent with matrigenic expression (Figure 2B). The two genes consistent with matrigenic expression were also consistent with patrigenic expression, as the predictions converged at several points (Figure 1).

Both *Major Royal Jelly Protein 3* and *Histone-lysine N-methyltransferase SETMAR* showed expression consistent with all six queens being homozygous for the focal allele, while the predictions from *SI RNA-binding protein* were unclear (Figure 2). We therefore evaluated the queens' genotypes by assessing allelic proportions within pools of their haploid sons, which closely approximate the diploid maternal genome for a large enough ( $n = 50$ ) sample of sons. This test revealed significant mismatches for all three genes, as the pooled males were often intermediate to SNP-allele frequency expectations of 0 or 1 for queen homozygosity or 0.5 for heterozygosity (Figure 3A). We therefore tested a number of individual haploid males, which were also found to often have genotypes inconsistent with the presence of only a single locus for the four focal genes (Figure 3B). Finally, qPCR of serially diluted DNA samples showed that the concentration of *SI RNA-binding protein* increased, as expected when this gene is present in only a single copy. However, the two remaining genes (*Histone-lysine N-methyltransferase SETMAR* and *Major Royal Jelly Protein 3*) increased at roughly twice this rate—suggesting that two copies of each of these genes were present in the genome (Figure 4).

## Discussion

Social Hymenoptera with a life-time unmated worker caste are predicted to utilise genomic imprinting in multiple situations of reproductive conflicts, although some of the theory was more confident (Queller, 2003) than other assessments (Kronauer, 2008). However, compared to mammals, there are few systematic searches for imprinted genes in social insects and all are restricted to either honeybees or bumblebees (Kocher et al., 2015, Galbraith et al., 2016, Lonsdale

et al., 2017, Smith et al., 2020, Marshall et al., 2020), predominantly due to the difficulty of maintaining and breeding other social insect species under laboratory conditions. Here, we developed a general method to search for imprinted genes in social hymenopteran species without requiring reciprocal crosses, and we demonstrated the utility of this technique using the leaf-cutting ant *Acromyrmex echinator*. This highlighted a number of potentially imprinted genes, but their verification failed to confirm that any of these were in fact imprinted. We evaluate these findings in detail below and conclude that our screening method should be sound, suggesting that validations remain crucial and that high-quality reference genomes are required to achieve conclusive results that avoid false positives.

Complications arose when we confronted our model with data obtained from independent RNA/DNA sequencing. We admit that the use of RNA-sequencing to identify imprinted genes is vulnerable to false-positives, low power and high variation (Wang and Clark, 2014). Experimental designs that use reciprocal crosses are likely to offer more powerful relative contrasts than our approach, as they rely on comparing differences between allelic RNA ratios, while our study engaged in the more difficult task of detecting differences between DNA and RNA proportions directly. The approach that we adopted here has its own challenges because we may not observe ASE at the colony-level even with perfect mono-allelic expression, because unimprinted and imprinted predictions overlap (where lines cross in Figure 1). We may also not observe large differences between DNA and RNA allele frequencies under mono-allelic expression compared to bi-allelic expression if the same SNP allele is present in both the queen and the males she mated with. This could be particularly problematic in species with high queen-mating frequencies (e.g. honeybees, or leaf-cutting ants as used here), as the likelihood that both alleles are present in the pool of fathers increases with the number of males contributing to a queen's sperm store.

Although these issues can be ameliorated by sufficient biological replication and sequencing depth, they reduce the difference between imprinted and unimprinted expectations, making them harder to distinguish statistically. Finally, our estimates of DNA and RNA proportions produced from the sequencing data were assumed to follow binomial distributions to construct the confidence intervals—a tenuous assumption due to biases in sequencing technologies (Wang and Clark, 2014).

Further, careful choice of samples is required. While the type of data from Li et al. (2014) were ideal for testing differences in SNP-allele frequencies between paired DNA and RNA samples, the tissues sampled for RNA extractions, i.e. adult heads, were perhaps not ideal for identifying imprinting. The tests here assume that imprinting is consistent across all individuals and samples, as with other imprinting studies (Wang and Clark, 2014), but we may not necessarily expect that imprinting is consistent across castes. For example, conflict over caste determination is already resolved in adults, while conflict over selfish worker male production may not be present in the smallest workers who do not tend to lay eggs (Dijkstra et al., 2005). In mammals, imprinting patterns can differ among tissues and different life stages (Wang et al., 2008), so an analogous pattern may also occur among castes in social insects. Excluding the gynes (who are not a life-time unmated helper caste) increased the number of putatively imprinted genes, which is consistent with caste-specific signatures of possible imprinting, but was expected due to the necessarily less stringent criteria. The data by Li et al. (2014) did not allow us to pursue this further because the method described here relies on multiple samples per colony to test for within-colony consistency. Samples in any future studies along the design developed here should therefore be carefully considered to avoid ambiguities of this kind: including multiple samples per colony per caste, for example, would greatly improve the power of this approach, and would enable searching for caste-specific effects.

Despite these caveats, we feel that the procedure we developed here has two compensating advantages. First, as already mentioned, our method can be applied to any haplodiploid social insect, rather than having to fulfil the much more stringent condition of reciprocal crosses being feasible. The alternative to the approach we describe would be to sequence DNA and RNA from single individuals. This approach is increasingly feasible with the ever-shrinking costs of high-throughput sequencing technology, but is not without challenges either. Distinguishing between ASE that arises because of parental effects versus random monoallelic expression would still require rather many samples. Further, extracting sufficient DNA and RNA from individual social insects workers is possible, but challenging if one wishes to avoid PCR-based amplification steps during library preparation that could bias allelic ratios. This would not have been a problem for the large workers of *A. echinator*, but might preclude easy application in ants with much smaller workers, not to mention for eggs or larvae. The approach we describe solves these problems because a sufficiently large pool of individuals would hide any random ASE (Wang and Clark,

2014) and thus leave the investigator with ample material for sample preparation provided colonies are big enough. Second, the method used here requires only a single sample to show significant ASE for a specific locus to qualify for the kind of combinatorial tests, rather than requiring that many samples differ from unimprinted expectations to reach significance, as is typical in many other imprinting studies (Wang and Clark, 2014). Thus, while our initial identification of ASE suffers from a high false discovery rate, our combinatorial follow-up approach and subsequent validation was more stringent—relying on specific, quantitative predictions that all samples had to fulfil before a gene could be considered as potentially imprinted. The method described here thus casts a wide net, but narrows this to a stricter subset of promising loci during the validation process.

### **Application to *Acromyrmex echinator***

We identified 517 SNPs that showed significant ASE in at least one caste sample, which qualified these loci for testing for consistency with imprinting predictions in all 9 samples (3 female castes in 3 colonies). Testing the RNA-DNA ratios for consistency with imprinting identified 59 SNPs consistent with patrigenic biased expression (at  $\alpha < 0.2$ ), and 55 SNPs consistent with matrigenic biased expression (at  $\alpha > 0.8$ ); overlap between predictions meant that a total of 92 SNPs were consistent with biased ASE under imprinting. Approximately 2/3 (63 of 92) of the putatively imprinted SNPs we identified were consistent with complete mono-allelic expression, and this percentage was almost twice as high for patrigenes (ca. 80%, 47 of 59) than matrigenes (ca. 45 %, 25 of 45). That fewer SNPs were consistent with exclusively matrigenic expression, particularly when demanding 100% biased expression, may reflect a difference in test stringency, rather than biology: matrigenic predictions depend on queen genotype rather than colony DNA proportions (slopes of 0 in Figure 1A), while patrigenic RNA predictions are proportional to DNA proportions (slopes of 2 in Figure 1B) and have confidence intervals twice that found in the DNA. A bias towards patrigenic genes has, however, previously been predicted in polyandrous social insect species (Queller, 2003), and has also been reported in the honeybees (Galbraith et al., 2016), which identified more genes consistent with patrigenic expression than found here, suggesting that our criteria are not too lenient.

We selected four putatively imprinted SNPs for further testing: two that were consistent with patrigenic expression (*Major Royal Jelly Protein 3* and *S1 RNA-binding domain containing*

protein) and two consistent with matrigenic imprinting (*Histone-lysine N-methyltransferase SETMAR* and *Vitellogenin 1*) for validation. These four loci were chosen based on how well they matched imprinting predictions (total difference between sequencing data and predictions, and feasibility of designing specific primers and probes), and therefore represented the strongest imprinting candidates the validation phase. While verification with ddPCR showed that none of these four selected genes is likely imprinted, it highlighted the utility of our validation method. The allelic proportions obtained from sequencing matched the values found by ddPCR in three of the four genes, but eliminated *Vitellogenin 1* from most further considerations (Figure 2). While the DNA and RNA allele frequencies of *S1 RNA-binding domain containing protein* produced by ddPCR were best explained as unimprinted, the remaining two genes *Major Royal Jelly Protein 3* and *Histone-lysine N-methyltransferase SETMAR* were consistent with 100% mono-allelic expression (Figure 2). Our model thus made clear predictions regarding maternal genotype, and predicted all queens to be homozygous for the focal allele. That result raised suspicion as being unlikely, and prompted validation of queen genotypes using pools of her haploid sons, which indeed showed intermediate genotypes that could only be explained by inferring that these genes were present in multiple copies (Figure 3A). For two genes we were able to confirm the presence of two copies via testing of individual males (Figure 3B) and by direct comparison to another gene known to be present in a single copy (Figure 4). Even in the highly unlikely scenario that some of the males collected were worker sons rather than queen sons, this would not affect the relative proportion of single and multiple copy genes within a colony as estimated here (Figure 4). Thus, it seems unlikely that these genes are imprinted, because the presence of two gene copies would produce imprinting-like-expression patterns if only one of the two genes were expressed in the tissue used for RNA extractions.

Rather than a weakness specific to this study, it seems likely that the presence of two gene copies highlights a weakness of all imprinting studies that do not explicitly consider this potential complication. While this possibility does diminish with increasing genome quality, there is no *a priori* reason to assume that the sequenced honey bee and bumblebee genomes are of much better quality than the *A. echinator* genome (Kocher et al., 2015, Galbraith et al., 2016, Lonsdale et al., 2017). Experimental designs using reciprocal crosses may avoid these issues because they directly target differences in RNA-sequences, rather than comparing RNA-sequences to DNA-sequences as we did (eg. Kocher et al., 2015, Galbraith et al., 2016, Smith et al., 2020), but we cannot



necessarily assume they are immune to duplicated gene complications. Because reciprocal crosses imply that putatively imprinted genes are expressed in different cytoplasmatic backgrounds, this design would also be liable to unexpected expression patterns should the target gene be present in two copies. Explicit testing for potential gene duplications thus seems to be a necessary step to include in social insect studies investigating the presence and possible impact of imprinted genes. Without these verifications, two of the four genes that we investigated would have been erroneously reported as being imprinted.

### **Acknowledgments.**

We are thankful to the Smithsonian Tropical Research Institute in Panama for the use of facilities and the Autoridad Nacional del Ambiente of Panama for issuing collection and export permits. We are also grateful to the anonymous reviewers for their comments that greatly improved this manuscript. This work was supported by the ERC Advanced Grant of JJB (grant number 323085).

### **Data availability**

Analyses in this article can be reproduced using the data and code provided by Howe et al. (2020).

### **REFERENCES**

- AGRESTI, A. & COULL, B. A. 1998. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52, 119-126.
- AMARASINGHE, H. E., CLAYTON, C. I. & MALLON, E. B. 2014. Methylation and worker reproduction in the bumble-bee (*Bombus terrestris*). *Proceedings of the Royal Society B-Biological Sciences*, 281.
- AMARASINGHE, H. E., TOGHILL, B. J., NATHANAEL, D. & MALLON, E. B. 2015. Allele specific expression in worker reproduction genes in the bumblebee *Bombus terrestris*. *PeerJ*, 3, e1079.
- BOOMSMA, J. J. & GRAFEN, A. 1991. Colony-level sex ratio selection in the eusocial Hymenoptera. *Journal of Evolutionary Biology*, 4, 383-407.
- BURT, A. & TRIVERS, R. 2006. *Genes in conflict: the biology of selfish genetic elements*, Harvard University Press.
- DIJKSTRA, M. B., NASH, D. R. & BOOMSMA, J. J. 2005. Self-restraint and sterility in workers of *Acromyrmex* and *Atta* leafcutter ants. *Insectes Sociaux*, 52, 67-76.

- DOBATA, S. & TSUJI, K. 2012. Intragenomic conflict over queen determination favours genomic imprinting in eusocial Hymenoptera. *Proceedings of the Royal Society B: Biological Sciences*, 279, 2553--2560.
- GALBRAITH, D. A., KOCHER, S. D., GLENN, T., ALBERT, I., HUNT, G. J., STRASSMANN, J. E., QUELLER, D. C. & GROZINGER, C. M. 2016. Testing the kinship theory of intragenomic conflict in honey bees (*Apis mellifera*). *Proc Natl Acad Sci U S A*, 113, 1020-5.
- GEHRING, M. & SATYAKI, P. R. 2017. Endosperm and imprinting, inextricably linked. *Plant Physiology*, 173, 143.
- GIBSON, J. D., ARECHAVALETA-VELASCO, M. E., TSURUDA, J. M. & HUNT, G. J. 2015. Biased allele expression and aggression in hybrid honeybees may be influenced by inappropriate nuclear-cytoplasmic signaling. *Frontiers in Genetics*, 6, 343.
- GRAFEN, A. 1985. A geometric view of relatedness. *Oxford surveys in evolutionary biology*, 2, 28-89.
- GUZMAN-NOVOA, E., HUNT, G. J., PAGE, R. E., URIBE-RUBIO, J. L., PRIETO-MERLOS, D. & BECERRA-GUZMAN, F. 2005. Paternal effects on the defensive behavior of honeybees. *Journal of Heredity*, 96, 376-380.
- HAIG, D. 1996. Placental hormones, genomic imprinting, and maternal-fetal communication. *Journal of Evolutionary Biology*, 9, 357--380.
- HAIG, D. 2000. The kinship theory of genomic imprinting. *Annual review of ecology and systematics*, 9--32.
- HAIG, D. 2004. Genomic imprinting and kinship: how good is the evidence? *Annual Review of Genetics*, 38, 553--585.
- HAIG, D. 2014. Coadaptation and conflict, misconception and muddle, in the evolution of genomic imprinting. *Heredity*, 113, 96--103.
- HAIG, D. & WESTOBY, M. 1989. Parent-specific gene expression and the triploid endosperm. *The American Naturalist*, 134, 147-155.
- HAIG, D. & WESTOBY, M. 1991. Genomic imprinting in endosperm: its effect on seed development in crosses between species, and its implications for the evolution of apomixis. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 333, 1-13.
- HAMILTON, W. D. 1964a. The genetical evolution of social behaviour. I. *J Theor Biol*, 7, 1-16.

- HAMILTON, W. D. 1964b. The genetical evolution of social behaviour. II. *J Theor Biol*, 7, 17-52.
- HOWE, J., SCHIØTT, M., LI, Q., WANG, Z., ZHANG, G., BOOMSMA, J.J. (2020), A novel method for using RNA-seq data to identify imprinted genes in social Hymenoptera with multiply mated queens, v3, Dryad, Dataset, <https://doi.org/10.5061/dryad.8gtht76n0>
- KENT, W. J. 2002. BLAT—the BLAST-like alignment tool. *Genome Res*, 12, 656-64.
- KÖCHER, S. D., TSURUDA, J. M., GIBSON, J. D., EMORE, C. M., ARECHAVALETA-VELASCO, M. E., QUELLER, D. C., STRASSMANN, J. E., GROZINGER, C. M., GRIBSKOV, M. R., SAN MIGUEL, P., WESTERMAN, R. & HUNT, G. J. 2015. A search for parent-of-origin effects on honey bee gene expression. *G3 (Bethesda)*, 5, 1657-62.
- KRONAUER, D. J. C. 2008. Genomic imprinting and kinship in the social *Hymenoptera*: what are the predictions? *Journal of Theoretical Biology*, 254, 737--740.
- LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754-60.
- LI, Q., WANG, Z., LIAN, J., SCHIOTT, M., JIN, L., ZHANG, P., ZHANG, Y., NYGAARD, S., PENG, Z., ZHOU, Y., DENG, Y., ZHANG, W., BOOMSMA, J. J. & ZHANG, G. 2014. Caste-specific RNA editomes in the leaf-cutting ant *Acromyrmex echinatior*. *Nat Commun*, 5, 4943.
- LIBBRECHT, R. & KELLER, L. 2013. Genetic compatibility affects division of labor in the Argentine ant *Linepithema humile*. *Evolution*, 67, 517-24.
- LIBBRECHT, R., SCHWANDER, T. & KELLER, L. 2011. Genetic components to caste allocation in a multiple-queen ant species. *Evolution*, 65, 2907-15.
- LONSDALE, Z., LEE, K., KIRIAKIDU, M., AMARASINGHE, H., NATHANAEL, D., O'CONNOR, C. J. & MALLON, E. B. 2017. Allele specific expression and methylation in the bumblebee, *Bombus terrestris*. *PeerJ*, 5, e3798.
- MARSHALL, H., VAN ZWEDEN, J. S., VAN GEYSTELEN, A., BENAETS, K., WÄCKERS, F., MALLON, E. B. & WENSELEERS, T. 2020. Genome-wide search for parent-of-origin allele specific expression in *Bombus terrestris*. *bioRxiv*, 2020.01.17.909168.
- MOORE, T. & HAIG, D. 1991. Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet*, 7, 45-9.

- NYGAARD, S., ZHANG, G., SCHIOTT, M., LI, C., WURM, Y., HU, H., ZHOU, J., JI, L., QIU, F., RASMUSSEN, M., PAN, H., HAUSER, F., KROGH, A., GRIMMELIKHUIJZEN, C. J., WANG, J. & BOOMSMA, J. J. 2011. The genome of the leaf-cutting ant *Acromyrmex echinator* suggests key adaptations to advanced social life and fungus farming. *Genome Res*, 21, 1339-48.
- OLDROYD, B. P., ALLSOPP, M. H., ROTH, K. M., REMNANT, E. J., DREWELL, R. A. & BEEKMAN, M. 2014. A parent-of-origin effect on honeybee worker ovary size. *Proceedings of the Royal Society B-Biological Sciences*, 281.
- PEGORARO, M., MARSHALL, H., LONSDALE, Z. N. & MALLON, E. B. 2017. Do social insects support Haig's kin theory for the evolution of genomic imprinting? *Epigenetics*, 12, 725-742.
- QUELLER, D. C. 2003. Theory of genomic imprinting conflict in social insects. *Bmc Evolutionary Biology*, 3, 15.
- QUELLER, D. C. & STRASSMANN, J. E. 2002. The many selves of social insects. *Science*, 296, 311-313.
- REIK, W. & WALTER, J. 2001. Genomic imprinting: parental influence on the genome. *Nat Rev Genet*, 2, 21-32.
- SMITH, N. M. A., YAGOUND, B., REMNANT, E. J., FOSTER, C. S. P., BUCHMANN, G., ALLSOPP, M. H., KENT, C. F., ZAYED, A., ROSE, S. A., LO, K., ASHE, A., HARPUR, B. A., BEEKMAN, M. & OLDROYD, B. P. 2020. Paternally-biased gene expression follows kin-selected predictions in female honey bee embryos. *Mol Ecol*, 29, 1523-1533.
- SPENCER, H. G. & CLARK, A. G. 2014. Non-conflict theories for the evolution of genomic imprinting. *Heredity*, 113, 112--118.
- TRIVERS, R. L. & HARE, H. 1976. Haplodiploidy and the evolution of the social insect. *Science*, 191, 249.
- ÚBEDA, F. & GARDNER, A. 2010. A model for genomic imprinting in the social brain: juveniles. *Evolution*.
- ÚBEDA, F. & GARDNER, A. 2011. A model for genomic imprinting in the social brain: adults. *Evolution*, 65, 462--475.
- ÚBEDA, F. & GARDNER, A. 2012. A model for genomic imprinting in the social brain: elders. *Evolution*, 66, 1567--1581.

- UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B. C., REMM, M. & ROZEN, S. G. 2012. Primer3—new capabilities and interfaces. *Nucleic acids research*, 40, e115-e115.
- VILLESEN, P., MURAKAMI, T., SCHULTZ, T. R. & BOOMSMA, J. J. 2002. Identifying the transition between single and multiple mating of queens in fungus-growing ants. *Proc Biol Sci*, 269, 1541-8.
- WANG, X. & CLARK, A. G. 2014. Using next-generation RNA sequencing to identify imprinted genes. *Heredity*, 113, 156-166.
- WANG, X., SUN, Q., MCGRATH, S. D., MARDIS, E. R., SOLOWAY, P. D. & CLARK, A. G. 2008. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *Plos One*, 3.
- WANG, Z., LIAN, J., LI, Q., ZHANG, P., ZHOU, Y., ZHAN, X. & ZHANG, G. 2016. RES-Scanner: a software package for genome-wide identification of RNA-editing sites. *GigaScience* [Online], 5. Available: <http://europepmc.org/abstract/MED/27538485>  
<https://doi.org/10.1186/s13742-016-0143-4>  
<https://europepmc.org/articles/PMC4989487>  
<https://europepmc.org/articles/PMC4989487?pdf=render> [Accessed 2016/08//].
- WEST, S. A., GRIFFIN, A. S. & GARDNER, A. 2007. Evolutionary explanations for cooperation. *Curr Biol*, 17, R661-72.
- WILD, G. & WEST, S. A. 2009. Genomic imprinting and sex allocation. *The American Naturalist*, 173, E1--14.
- WITHROW, J. M. & TARPY, D. R. 2018. Cryptic “royal” subfamilies in honey bee (*Apis mellifera*) colonies. *PLOS ONE*, 13, e0199124.
- WOLFF, P., WEINHOFFER, I., SEGUIN, J., ROSZAK, P., BEISEL, C., DONOGHUE, M. T. A., SPILLANE, C., NORDBORG, M., REHMSMEIER, M. & KÖHLER, C. 2011. High-resolution analysis of parent-of-origin allelic expression in the *Arabidopsis* endosperm. *PLoS Genetics*, 7, e1002126.

## FIGURES

**Figure 1: Expected relationships between SNP frequencies in RNA and DNA from pooled samples of individuals collected from haplodiploid social insect colonies headed by one multiply inseminated queen under imprinted expression of a focal gene.** Without imprinting, DNA and RNA are equal (dashed lines). The coloured lines represent the relationship under, respectively, 100% matrigenic (left, horizontal lines) and 100% patrigenic expression (right, oblique lines). Line colour indicates queen genotype, as homozygous (green & blue) or heterozygous (red). The shaded areas indicate predictions where bias in allele-specific expression is high ( $\geq 80\%$ ) but not complete.

**Figure 2: Relationships between SNP-allele frequencies in RNA and DNA for four candidate imprinted genes.** The relationship between SNPs in the DNA and the RNA was obtained using either sequencing data from Li et al (2014) (A) or ddPCR (B) for each of the four candidate genes. Bars indicate either the 95% confidence intervals based on allele proportions from DNA and RNA sequencing (A), or error estimates from Poisson distributed templates among droplets in ddPCR reactions (B). The coloured lines and shaded areas represent ASE predictions under imprinting while dashed lines indicate unimprinted expectations as in Figure 1. Two of the selected genes showed expression patterns consistent with patrigenic expression (left, *Major Royal Jelly Protein 3*, and *S1 RNA-binding domain containing protein 1*), and two with matrigenic expression (right, *Histone-lysine N-methyltransferase SETMAR*, and *Vitellogenin 1*).

**Figure 3: SNP genotypes of mothers and sons and gene copy number assessments: data from queen sons failed to produce evidence for a single diploid queen DNA genotype.** (A) Pools of haploid males analysed by ddPCR gave SNP allelic ratios inconsistent with a single diploid genome (error bars as in Figure 2). Instead, we often obtained SNP-allele frequencies intermediate between expected homozygosity (0, 1) and heterozygosity (0.5) values (represented by the three horizontal dashed lines, colours indicate queen genotype as in Figure 2). The three genes tested were the same as in figure 2: *Major Royal Jelly Protein 3*, *S1 RNA-binding domain containing protein*, and *Histone-lysine N-methyltransferase SETMAR*. The RNA-seq data (Figure 2a) predicted maternal homozygosity for the focal SNP at all genes; the ddPCR data for *Major Royal Jelly Protein 3* and *Histone-lysine N-methyltransferase SETMAR* agreed with this (Figure 2b), but these predictions were not upheld when we analysed the mother queen genotypes from pools of

their haploid sons. **(B)** Our ddPCR analyses of SNP genotypes of individual queen sons were inconsistent with the presence of a single gene copy, i.e. we did not obtain a single copy locus except for *S1 RNA-binding domain containing protein*. The remaining genes (*Major Royal Jelly Protein 3*) showed SNP-allele frequencies among queen sons consistent with two gene copies within the haploid genome, and SNP allelic proportions for *Histone-lysine N-methyltransferase SETMAR* suggested non-specific binding and amplification in colony Ae322.

**Figure 4. Independent copy number assessment of three focal genes by direct measurement of their concentration compared to the concentration of a reference gene *TATA-Box Protein* with confirmed single copy status.** Dashed lines indicate the theoretical expectations for a single copy gene (1:1) and for a duplicated gene (2:1). Two genes (*Histone-lysine N-methyltransferase SETMAR* and *Major Royal Jelly Protein 3*) had concentrations consistent with the presence of two copies (in agreement with Figure 3 for *Major Royal Jelly Protein 3*, but not necessarily for *Histone-lysine N-methyltransferase SETMAR*). The remaining gene (*S1 RNA-binding domain containing protein*) had concentrations consistent with a single gene copy (in agreement with Figure 3B).

**Figure 1**

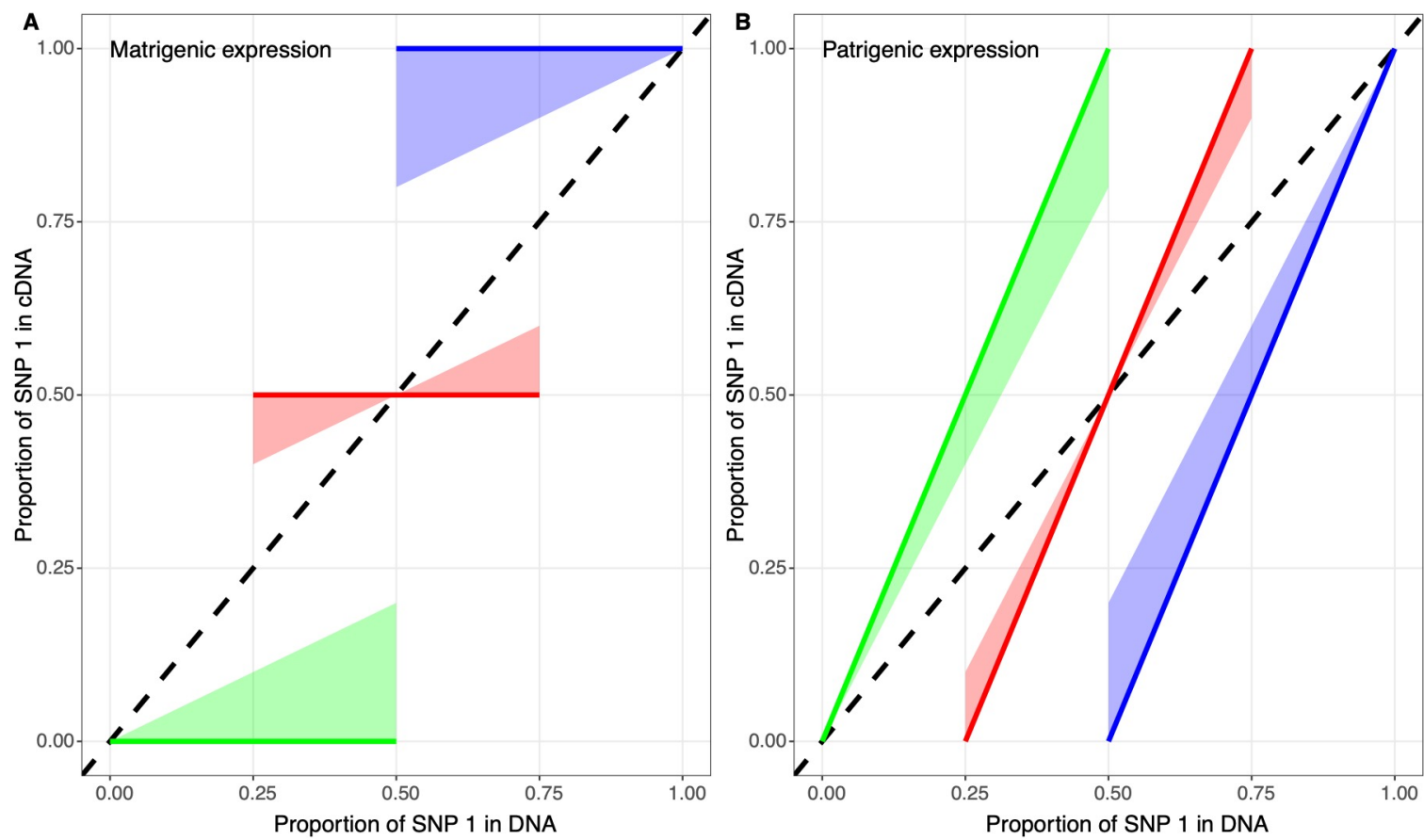




Figure 2

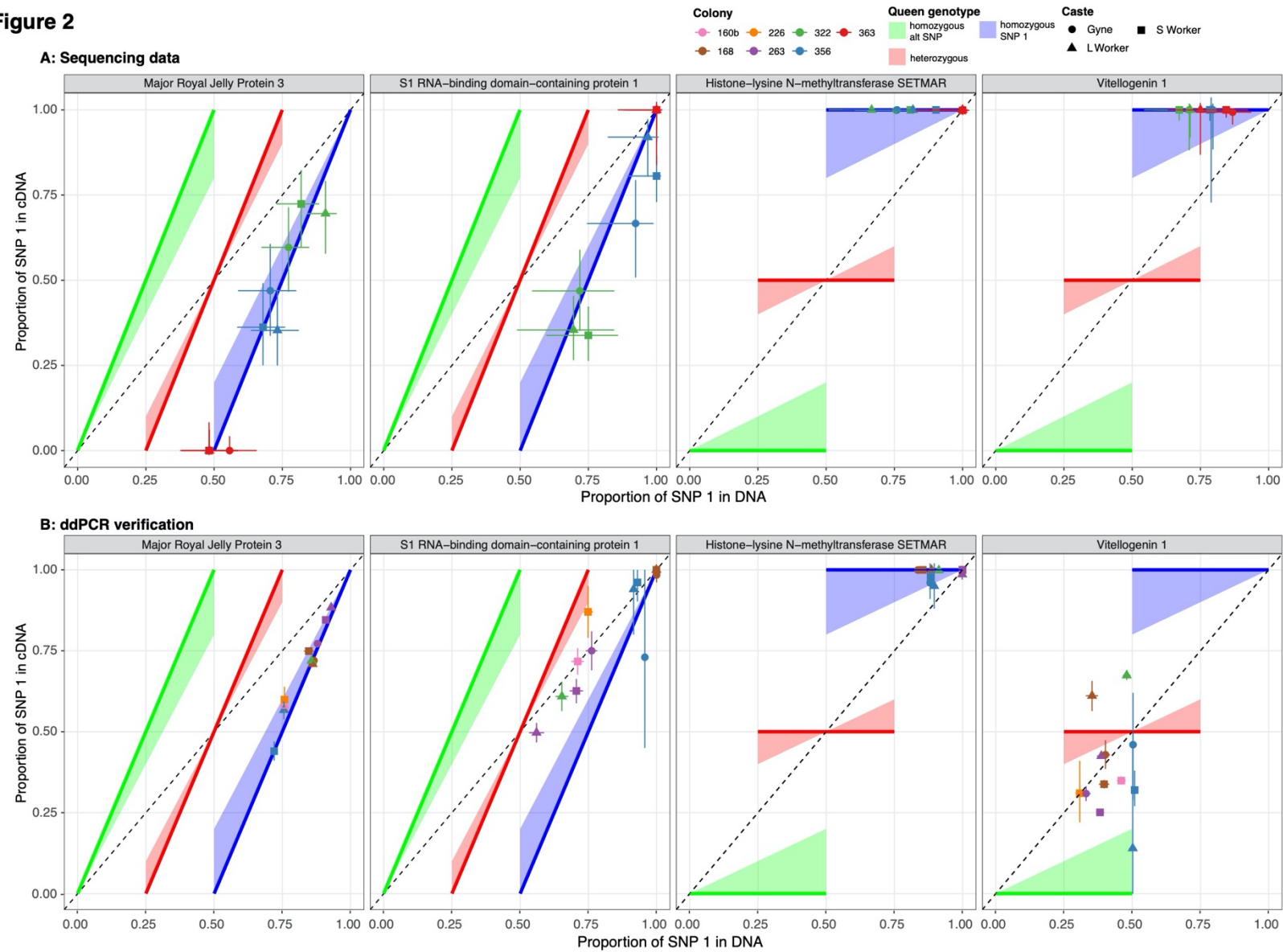
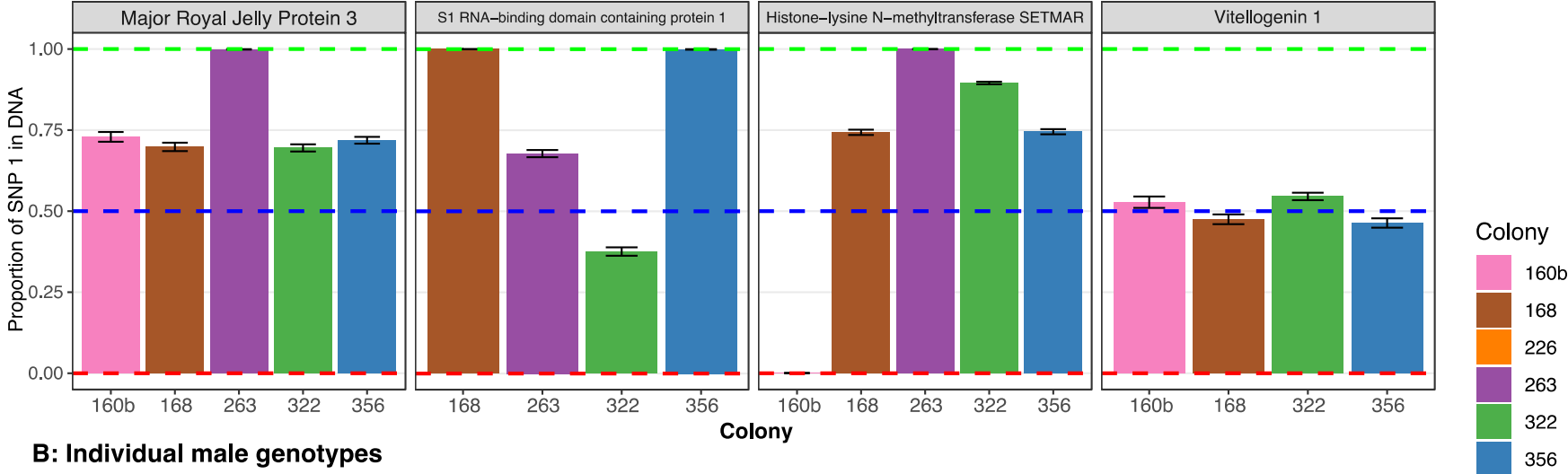


Figure 3

**A: Queen genotypes**



**B: Individual male genotypes**

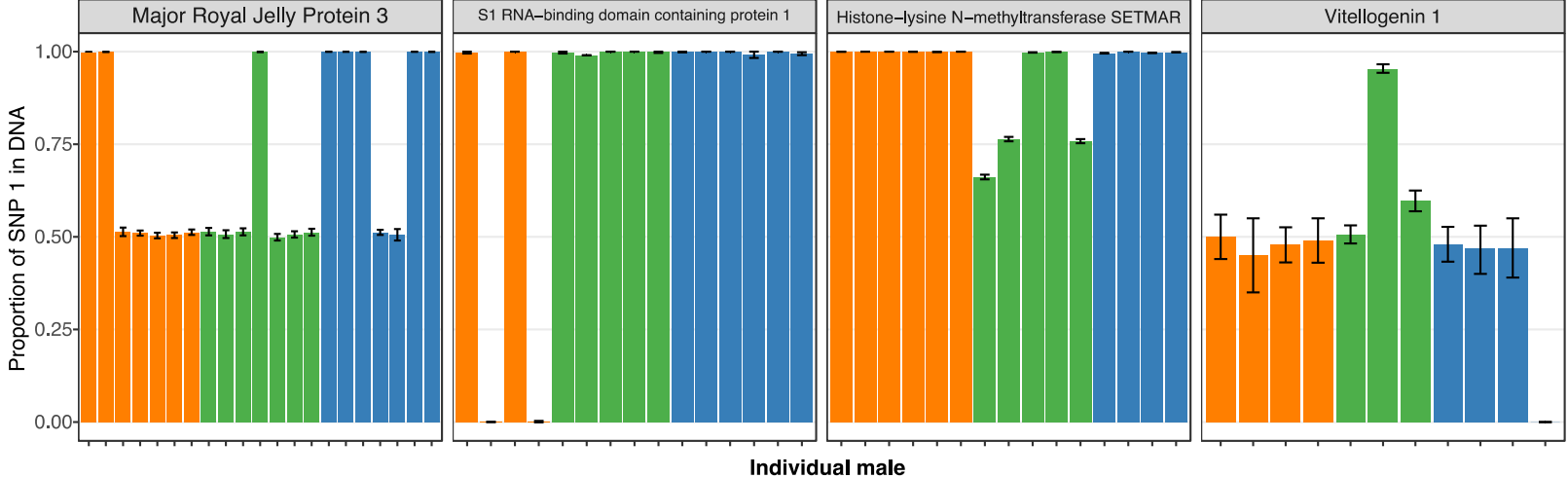


Figure 4

