Transcriptome sequencing reveals high isoform diversity in the ant *Formica exsecta*

Kishor Dhaygude[1,*], Kalevi Trontti[2,*], Jenni Paviala[1], Claire Morandin[1], Christopher Wheat[3], Liselotte Sundström[1,4] and Heikki Helanterä[1,4]

[1] Centre of Excellence in Biological Interactions, Department of Biosciences, University of Helsinki, Helsinki, Finland
[2] Department of Biosciences, Neurogenomics Laboratory, University of Helsinki, Helsinki, Finland
[3] Department of Zoology Ecology, Stockholm University, Stockholm, Sweden
[4] Tvärminne Zoological Station, University of Helsinki, Hanko, Finland
[*] These authors contributed equally to this work.

ABSTRACT

Transcriptome resources for social insects have the potential to provide new insight into polyphenism, i.e., how divergent phenotypes arise from the same genome. Here we present a transcriptome based on paired-end RNA sequencing data for the ant *Formica exsecta* (Formicidae, Hymenoptera). The RNA sequencing libraries were constructed from samples of several life stages of both sexes and female castes of queens and workers, in order to maximize representation of expressed genes. We first compare the performance of common assembly and scaffolding software (Trinity, Velvet-Oases, and SOAPdenovo-trans), in producing *de novo* assemblies. Second, we annotate the resulting expressed contigs to the currently published genomes of ants, and other insects, including the honeybee, to filter genes that have annotation evidence of being true genes. Our pipeline resulted in a final assembly of altogether 39,262 mRNA transcripts, with an average coverage of >300X, belonging to 17,496 unique genes with annotation in the related ant species. From these genes, 536 genes were unique to one caste or sex only, highlighting the importance of comprehensive sampling. Our final assembly also showed expression of several splice variants in 6,975 genes, and we show that accounting for splice variants affects the outcome of downstream analyses such as gene ontologies. Our transcriptome provides an outstanding resource for future genetic studies on *F. exsecta* and other ant species, and the presented transcriptome assembly can be adapted to any non-model species that has genomic resources available from a related taxon.

Subjects Bioinformatics, Computational Biology, Entomology, Evolutionary Studies, Genomics
Keywords Pool-seq, RNA-sequencing, Ants, Transcriptome *de novo* assembly, Hymenoptera, Transcriptomics

Submitted 14 April 2017
Accepted 17 October 2017
Published 21 November 2017

Corresponding author
Kishor Dhaygude,
kishor.dhaygude@helsinki.fi

Academic editor
Dick de Ridder

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.3998**

INTRODUCTION

Phenotypic variation can arise via differences in gene sequences or patterns of gene expression (Carroll, 2008; Simpson, Sword & Lo, 2011). Adaptive polyphenism is one of the most dramatic examples of how variation in gene expression can be translated into alternative phenotypes, the castes of social Hymenoptera (ants, social bees and social

How to cite this article Dhaygude et al. (2017), Transcriptome sequencing reveals high isoform diversity in the ant *Formica exsecta*. **PeerJ 5:e3998; DOI 10.7717/peerj.3998**

wasps) being a prime example (Gräff et al., 2007; Bonasio et al., 2010; Hunt et al., 2011; Colgan et al., 2011). The behavioural, physiological, and morphological differentiation between workers and reproductive queens, underlie the adaptive radiations and ecological importance of social Hymenoptera, especially ants (Wilson & Hölldobler, 2005). A diploid, fertilized hymenopteran egg is totipotent, i.e., has the genetic prospects to develop into a queen or a worker, and the developmental trajectory the egg takes among these alternatives is, with a few exceptions (Helms Cahan & Keller, 2003; Pearcy et al., 2004; Fournier et al., 2005), directly influenced by the nutrition and rearing conditions provided by workers (Schwander et al., 2010). These conditions presumably launch a cascade of differential gene expression of a few genes, with large pleiotropic effects at early larval instars causing them to develop into different female castes (Lattorff & Moritz, 2013). Importantly, caste determination is also controlled by epigenetic mechanisms that are regulated by the social environment, e.g., workers (Gräff et al., 2007; Weil, Korb & Rehli, 2009; Glastad et al., 2011; Zwier et al., 2012; Simola et al., 2013; Simola et al., 2016; Welch & Lister, 2014; Bonasio, 2014; Alvarado et al., 2015; Ashby et al., 2016).

Genetic causes and consequences of caste polyphenism have been studied mostly using preselected candidate genes (Abouheif, 2002; Hunt & Goodisman, 2010; Shbailat, Khila & Abouheif, 2010; Feldmeyer, Elsner & Foitzik, 2014), or based on phylogenetically limited comparisons of distantly related taxa (Hunt & Goodisman, 2010; Hunt et al., 2010; Ometto et al., 2011; Berens, Hunt & Toth, 2015). Current genomic resources of social insects are available for important pollinators such as honey bees and bumble bees (Colgan et al., 2011; Kocher et al., 2013; Elsik et al., 2014; Elsik et al., 2016; Kapheim et al., 2015; Sadd et al., 2015), and primitively eusocial paper wasps (Sumner, Pereboom & Jordan, 2006; Ferreira et al., 2013). For ants, genome sequences are currently available for 13, and transcriptomes for 23, species in the Fourmidable database (Wurm et al., 2009), including well studied species such as the invasive argentine ant *(Linepithema humile)*, the fire ant *(Solenopsis invicta),* and the leaf cutting ant species *(Acromyrmex echinatior* and *Atta cephalotes* (Bonasio et al., 2010; Suen et al., 2011; Smith et al., 2011a; Smith et al., 2011b; Wurm et al., 2011; Nygaard et al., 2011; Gadau et al., 2012; Oxley et al., 2014; Schrader et al., 2014; Konorov et al., 2017). However, these data cover only a fraction of the genetic richness of the more than 11,000 described ant species, and over 100 million years of evolution (Wilson & Hölldobler, 2005). More genomic resources are therefore required for phylogenetically informative comparisons that thoroughly assess the genomic consequences of sociality.

RNA-sequencing (high throughput sequencing of transcriptomes) is a relatively inexpensive way to rapidly sequence the coding and expressed genes of a species (Vera et al., 2008). However, there are technical challenges in using the resulting transcriptome assembly (TA) data from non-model organisms in downstream applications in a robust manner, posing problems for the kind of comparative work outlined above. First, *de novo* transcriptome libraries usually have contaminations with exogenous RNA from e.g., the microbial flora of the species (Bonfert et al., 2013), as demonstrated by a meta-transcriptome of our study species (Johansson et al., 2013). Second, TA's include transcripts that fail to be annotated, and align poorly to even moderately related species. Such unannotated contigs (contiguous consensus sequences that are derived from collections of overlapping reads in the assembly process) may include orphan genes, or variable, and less known regulatory non-coding RNA (*Wissler et al., 2013*). However, unannotated contigs may also arise from assembly errors, intron retention, or non-functional transcripts, such as pseudogenes (*Graveley et al., 2011*). As long as the true status of these remains unsolved, removing them from the TA increases its accuracy, and facilitates species comparisons, yet this process also leads to loss of information. Third, unique loci in the genome may be represented in a TA by dozens of predicted isoforms, which typically are not assembled completely. Given this, and the potential for overprediction of isoforms (i.e., false isoforms), care must be taken to properly use the TA to obtain unbiased read counts for each gene (*Tulin et al., 2013*). One approach is to identify likely gene-isoform relationships, and sum the reads mapping to each of these fragments and isoforms, that are inferred to come from the same locus (*Hornett & Wheat, 2012*).

Here we present a transcriptome of the ant *Formica exsecta* (Formicidae; Hymenoptera), a common species throughout the Palearctic region. It has morphologically well-separated queen and worker castes, and forms perennial nests of some thousands of workers (*Brown, Keller & Sundström, 2002*; *Vitikainen, Haag-Liautard & Sundström, 2011*; *Vitikainen, Haag-Liautard & Sundström, 2015*). Genetic and social nest structure varies greatly, and each nest can have one or several queens that have mated with one or many males, which makes it a model species for kin conflicts (*Sundström, Chapuisat & Keller, 1996*; *Brown, Liautard & Keller, 2003*; *Kummerli & Keller, 2007*), social recognition (*Martin et al., 2008*; *Martin et al., 2009*; *Van Zweden et al., 2011*), and causes and consequences of inbreeding (*Sundström, Keller & Chapuisat, 2003*; *Haag-Liautard et al., 2009*; *Vitikainen, Haag-Liautard & Sundström, 2011*; *Vitikainen, Haag-Liautard & Sundström, 2015*).

We characterise the transcriptome of *Formica exsecta,* including different sexes, life stages, and castes to obtain a good representation of expressed genes with their possible isoforms, potentially useful as a genomic resource for future studies on ants, as well as other insects. The transcriptome was generated from paired-end RNA-seq on seven libraries: cocoons of workers and queens, young workers and queens newly emerged from cocoons, overwintered adult workers and queens, and a pooled library of males (cocoons and young males newly emerged from cocoons). In order to avoid the TA-related challenges outlined above, we took additional steps of documenting biological evidence for the discovered contigs, by validating them with published ant genomes and genomes of other insects, and clustering the predicted isoforms to unique genes. We also show that different life stages/castes of ants can have different genes expressed. This provides more comprehensive information on the potential number of expressed genes, and their isoform numbers in ants. Furthermore, we provide a commented guideline on the comparison of analysis procedures, i.e., how to obtain overall expression profiling without a reference genome.

METHODS

Sampling and library sequencing

Samples were collected from six localities within a range of 50 km on the Hanko Peninsula, and the islands outside Tvärminne Zoological Station, SW Finland, between late April and July 2011 (Bioproject: PRJNA213662, Biosample: SAMN02297446–SAMN02297452 *Johansson et al., 2013*; *Morandin et al., 2014*; *Morandin et al., 2015*). Old reproductive queens and old overwintered workers were sampled directly from field colonies in late April and early May, when the queens aggregate at the colony surface. Old males are not available for analyses as they do not live past the mating season. Queen, male, and worker cocoons were collected from the same populations in late June and early July the same year, at an age when sex and caste can be visually assessed. The cocoons consisted of three stages: young (white cuticle and eyes), intermediate (white cuticle but pigmented eyes), and old (pigmented cuticle and eyes). Young males and queens were sampled prior to mating when they appeared at the surface of the nest, and were about to leave for their nuptial flight. Young workers were sampled within a week of eclosion, while still pale, compared to old workers. All samples were frozen directly after collections, and placed in −80 °C awaiting RNA extraction.

The samples were cleaned from visible exogenous material under a preparation microscope, and the total RNA was extracted from whole-ground ants individually in TriSure (Bioline, London, UK), following the manufacturers protocol. RNA quality was determined by assessing the integrity of ribosomal RNA with BioAnalyzer total RNA kit (Agilent, Santa Clara, CA, USA), and denaturing agarose electrophoresis. Subsequently each sample was pooled into seven RNA libraries (Table 1), so that total RNA of each sample had equal representation in the pool. The exception to this was the library of old workers, which could be equalized only between locations, due to the low quality of many extractions derived from old workers. The read data, and the details on RNA pools and sampling locations used here, correspond to the BGI-sequenced libraries in *Johansson et al. (2013)*. In short, altogether 105 individuals from 56 colonies were included in the libraries, and seven RNA libraries were constructed to cover sexes, female castes, and life stages; the paired-end sequencing libraries were constructed and sequenced by the Beijing Genomic Institute BGI (China), according to the provider's pipeline. The total RNA pools were DNAse-treated and selected for mRNA using poly-A-tail selection, and subsequently fragmented. Approximately 200 base insert length fragments were selected for library construction. Paired-end sequencing was conducted on Illumina HighSeq 2000 platform, 90 bases paired-end reads at Beijing Genomics Institute (BGI Illumina, Inc.). The raw reads of the transcriptome are available on GenBank (https://www.ncbi.nlm.nih.gov/sra/) under Bioproject ID PRJNA213662, SRA accession numbers: SRR945908, SRR945909, SRR945910, SRR945911, SRR945912, SRR945913, and SRR945914.

Transcriptome assemblies

The workflow of the assembly process is outlined in Fig. 1. Before the assembly process, we combined multiple raw data files generated from different sexes, life stages, and castes (Table 1). Altogether the combined raw data contained RNA expression profiling data from 105 individuals. To assess the quality of each sequencing library, quality checks were performed separately without combining raw data. The quality of raw reads was assessed separately for forward and reverse reads with FastQC (*Andrews, 2010*). Forward and reverse reads were trimmed with the FastX toolkit (Version 0.0.13 *Hannon, 2010*) to equal length

Table 1 **RNASeq sequencing libraries.** Caste, age, and the number of individuals pooled in each of the 7 libraries sequenced by BGI.

Caste	Age	# Individuals	# Colonies	Sequencing library
Queen	Old overwintered	4	4	Library 1- Old Queen
	New	8	4	Library 2- New Queen
	Old cocoon	6	3	Library 3- Queen Cocoon
	Intermediate cocoon	6	3	
	Young cocoon	6	3	
Worker	Old overwintered	30	8	Library 4- Old Worker
	New	15	6	Library 5- New Worker
	Old cocoon	6	3	Library 6- Worker Cocoon
	Intermediate cocoon	6	3	
	Young cocoon	6	3	
Male	Adult	3	3	Library 7- Male Mix
	Old cocoon	3	3	
	Intermediate cocoon	3	3	
	Young cocoon	3	3	

of 76 bp by removing the last 14 bp at the 3′-ends. After trimming, all reads were free from Illumina adapter contamination and primer dimers, paired data were without any orphan reads, and were of high quality (average $\geq$ 20 Q for all base positions on Phred scale). This resulted in 322 million clean read pairs for the TA (raw read processing summary table in File S1, Table S1A). These trimmed reads were assembled in four different stages of TA (Initial-TA, Meta-TA, Evidence-TA, and Unigene-TA) (Fig. 1).

Initial-TA assembly

In order to increase our confidence in the quality of the data, the initial *de novo* transcriptome assemblies (Initial-TA) were carried out using three commonly used transcriptome assembly software: Trinity (release 2012-05-18 (*Grabherr et al., 2011*; *Haas et al., 2013*), Velvet-Oases (version 0.2.06 (*Schulz et al., 2012*)), and SOAPdenovo-trans (version 2011-11-12 (*Xie et al., 2014*)). Each assembler was run with a range of k-mer values (Trinity k-mers: 21 to 31, Velvet-Oases k-mers: 51 to 75 and SOAPdenovo-trans k-mers: 21 to 71 File S1, Tables S1B–S1D), within the applicable range of the software. The minimum accepted contig length was set to 100. Of the three assemblers, we selected as the best TA assembler the method that covered the greatest proportion of orthologous genes with high coverage (>90% alignment coverage) in the genome of *Camponotus floridanus*. These are applicable not only to the assessment of transcriptome assemblies, but also annotated gene sets. For this comparison, we first constructed a blast database of *Camponotus floridanus* predicted proteins, CflorPP90, containing 15,977 proteins after removing redundant sequences from the original sets (17,064 proteins). Redundant sequences were removed by clustering *Camponotus floridanus* proteins, such that each cluster of similar sequences, with at least 90% sequence identity and 90% overlap, were represented by the longest sequence. This was done using in-house scripts (*Wheat, 2017*). Second, we aligned the Initial-TA contigs of each assembler (different k-mer settings) to

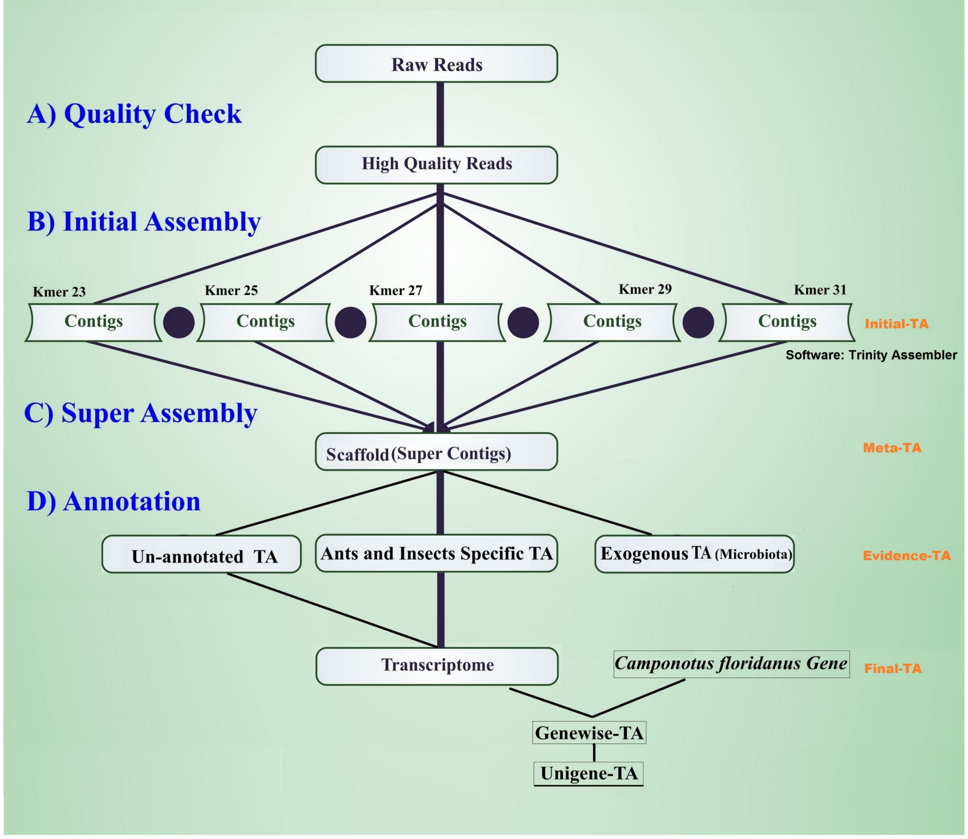

Figure 1 **Workflow for the transcriptome assembly.** High quality reads were assembled in contigs using five different k-mer settings (initial assembly), and merged (super assembly). The resulting contigs were investigated for evidence of being true species genes. False isoforms due to SNP variation, and sequencing errors were removed by self-alignment, and the remaining contigs with isoforms were assigned to genes using alignment to proteins of related species.

Full-size DOI: 10.7717/peerj.3998/fig-1

the predicted protein set CflorPP90 using BLASTx (ncbi-blast-2.2.2), with the minimum requirement of 50% amino acid identity, and an alignment length of at least 50 amino acids. Finally, we also recorded the number and length distributions of contigs arising from each assembly method (average contig length and N50), to compare the behaviour of different assemblers at different settings, but these values were not used as an assembly quality proxy (*Hornett & Wheat, 2012*; *O'Neil & Emrich, 2013*). We also made separate Trinity assemblies for each caste (male, queen, worker; Table 1.) with the default k-mer, and defined them as Male-TA, Queen-TA and Worker-TA. These caste specific assemblies were used in further evaluations of the final transcriptome assembly (Final-TA) by running BUSCO (*Simão et al., 2015*) assessment with Hymenoptera lineage specific orthologous gene sets.

Meta-TA assembly

The next assembly version (Meta-TA) was produced by combining the contigs that were obtained from all k-mer runs of the Initial-TAs from the Trinity assembler. This was done in order to assemble as many transcripts as possible, as short k-mer values are typically more sensitive than longer ones in detecting rare transcripts. Hence, short k-mers tend to produce more transcripts than long k-mers, whereas long k-mers typically produce fewer, but longer, intact contigs (*Surget-Groba & Montoya-Burgos, 2010*). We used Vmatch (*Abouelhoda, Kurtz & Ohlebusch, 2002*) to generate non-redundant sets of transcripts by combining contigs from different k-mer specific trinity assemblies, using default parameters, and by setting the minimum overlap for two contigs to collapse to be at least 100 bases, with a sequence identity of at least 95%. During scaffolding, contigs with very small differences were combined into a consensus sequence. This unavoidably removes some true variation, but we assume most of these isomorphs are artefacts, resulting from SNP and indel variation in the RNA of the pooled individuals, and from sequencing and assembly errors (*Nakamura et al., 2011*).

Evidence-TA assembly

To select TA contigs with evidence of being ant or other insect genes, we compared the Meta-TA contigs against non-redundant (NR) protein datasets from the NCBI database (Updated 2015), and swissprot databases using Blast2Go (BLASTx, *e*-value threshold 10^{-3} (*Conesa et al., 2005*)). This blast outcome was used to retrieve annotation, gene ontology terms (GO), enzyme annotation, and information of the protein family. Contigs annotated with known micro-organisms, and probable parasites such as viruses and fungi, were first filtered out from the data as exogenous material, as described in *Johansson et al. (2013)*, and this exogenous material has now been made available in NCBI database (Bioproject: PRJNA412141). The remaining contigs were then aligned (BLASTx, *e*-value threshold 10^{-3}) to (a) the predicted gene sets of published ant genomes available on the Fourmidable database, including the closest species such as *Camponotus floridanus*, *Lasius niger,* as well as more distantly related species such as *Cerapachys biroi*, *Linepithema humile*, *Solenopsis invicta*, *Vollenhovia emeryi*, *Wasmannia auropunctata*, *Pogonomyrmex barbatus*, *Monomorium pharaonis*, *Harpegnathos saltator*, *Acromyrmex echinatior* and *Atta cephalotes* (*Wurm et al., 2011*; *Wurm et al., 2009*; *Bonasio et al., 2010*; *Suen et al., 2011*; *Smith et al., 2011a*; *Smith et al., 2011b*; *Nygaard et al., 2011*; *Gadau et al., 2012*; *Oxley et al., 2014*; *Schrader et al., 2014*; *Konorov et al., 2017*), (b) the honey bee genome (*Weinstock et al., 2006*), (c) the predicted gene sets of *Nasonia vitripennis*, *Tribolium castaneum* (*Wurm et al., 2009*), and *Drosophila melanogaster* (*Gramates et al., 2017*), (d) as well as non-redundant (NR) protein datasets available in the NCBI database (Updated 2015, Updated 2017) for insect species. A minimum requirement of 70% amino acid identity, and at least 100 bases (33 amino acids) alignment length was used. The *L. niger* genome has been added with the NCBI 2017 database update. Contigs finding a match in these data sets were considered as true expressed genes, and form the TA supported by biological evidence (Evidence-TA assembly). This approach is conservative, as choosing only genes supported by annotation evidence may exclude genes with very high divergent sequence evolution, and hence possibly some orphan genes (*Tautz & Domazet-Lošo, 2011*).

Unigene-TA assembly

Evidence-TA contigs were aligned to the genome of *C. floridanus* using GenomeThreader (*Gremme et al., 2005*) to obtain information on gene structures (number and location of exons) of assembled transcripts, and to estimate the total number of genes and isoforms included in the Evidence-TA. *C. floridanus* diverged from *F. exsecta* approximately 80 million years ago (*Moreau et al., 2006*). In order to resolve the gene structure, all spliced alignments of transcripts were combined to generate a consensus alignment, which also helped to uncover alternative splicing. Isoform contigs were grouped to the *C. floridanus* genes by the exon(s), to which they show the greatest similarity. Using the genome of closely related species, instead of self-BLAST, also allows joining non-overlapping contigs to the reference genes (Perl Script present in File S2). We also manually validated 300 probable isoforms to assess the quality of isoform prediction obtained with this workflow (Fig. 2). After this, the longest isoform identified by GenomeThreader was chosen for inclusion in the final TA containing only unique and non-redundant contigs, supported by biological evidence. These can be considered as expressed sequence tags (ESTs) of *F. exsecta* (Unigene-TA). The longest isoform was chosen, as it is likely to contain all or most of the gene exons and is, with some caution such as recent gene duplications, comparable to a predicted codon set extracted from a genome, and can be used in a similar manner in downstream applications. Genes with very high DNA sequence divergence rates may, however, have been lost during this process. The Unigene-TA transcripts were finally compared to Protein databases (NCBI, swissprot) using the blast2go tool (*Conesa et al., 2005*) for annotations of gene ontology.

RESULTS AND DISCUSSION

Assemblies

In total, we obtained 322 M high quality reads. Based on ENCODE consortium recommendations, up to 100 million reads may be needed to successfully profile gene expression, with correctly detected isoforms in human data (*The ENCODE Consortium, 2011*). Since NGS technologies produce sequence data as short reads, and have a higher error rate (0.1–1%), a higher depth of sequencing is required for *denovo* assembly. Nevertheless, the optimal sequencing depth again depends on the aims of the experiment, non-uniformity of coverage in RNASeq, and sequencing error rate. For non-model organisms without a reference genome, the number of reads required should be scaled based on genome size, number of genes, and the total number of protein-coding genes. As the estimated number of the protein coding genes is somewhat smaller in ants than in humans (17,220 genes *Oxley et al., 2014*) compared to 19,836 in humans (GENCODE release 27, *Harrow et al., 2012*, and the total genome size much smaller (*Tsutsui et al., 2008*; *Harrow et al., 2012*), and our RNA-Seq data was obtained from different life stages and castes, we can therefore expect a reasonably high coverage, missing only transcripts with very low expression levels.

The transcriptome assemblers reported total assembly lengths of initial assemblies of up to 85–245 million bases (Mb), depending on software, with the maximum number of contigs generally between ca. 200 k and 300 k (File S1, Tables S1B–S1D). These are typical numbers for a *de novo* transcriptome assembly (*O'Neil & Emrich, 2013*; *Nakasugi et al.,*

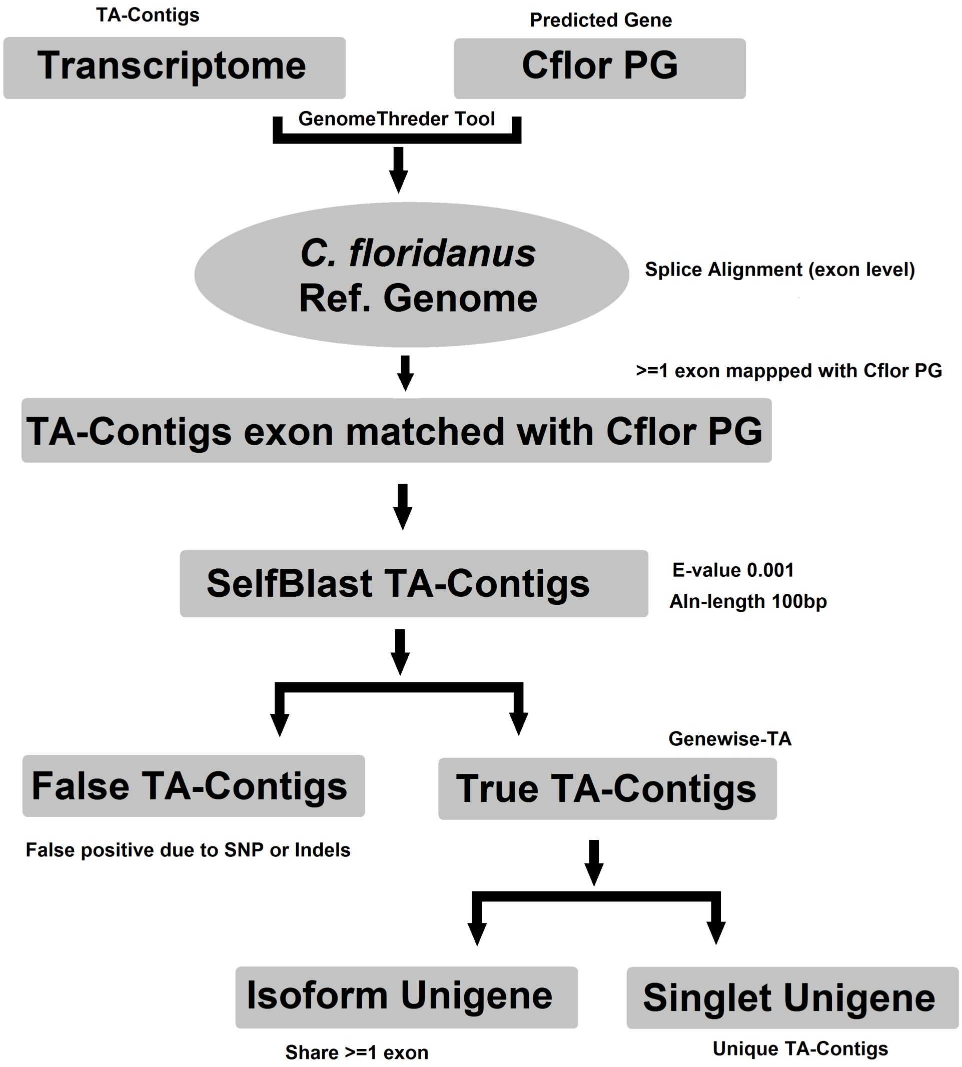

Figure 2 Workflow for compiling the Unigene-TA with isoform counts by contigs assignment to genes using alignment to proteins of related species.

Full-size DOI: 10.7717/peerj.3998/fig-2

2014; Rana et al., 2016). However, these are inevitably, at some level, inflated by structural variation, such as SNP and indels, given the use of pooled RNA from individuals of many populations and colonies, and non-source species contaminant sequences (Johansson et al., 2013). A reduced contig number in the initial assembly could potentially be achieved by using individuals with less genetic variability. However, this is not possible in *F. exsecta*, where most colonies produce reproductive individuals of one sex only (Sundström, Chapuisat & Keller, 1996), and sampling the old queens in the spring precludes sampling their offspring later in the season. The general expectation that short k-mers produce more, but shorter contigs than long k-mers, held for all tested software. SOAPdenovo-trans and Velvet-Oases gave the highest N50, and average contig lengths using relatively long k-mers between 50–61 (2,167/681.7 and 2,133/835.8 bases, respectively; File S1, Tables S1B and S1C). Conversely, Trinity gave the highest N50 and average contig length with a shorter k-mer size $K = 31$ (2,863/2,032 bases, respectively; File S1, Table S1D). Another notable difference between the Initial-TA assemblies of the three assemblers resulted from reported non-ATGC characters (N's), which formed up to 9.45% in the SOAPdenovoTrans assemblies and 0.017% of the Velvet-Oases assemblies, whereas the Trinity assemblies do not report any unknown bases (File S1, Tables S1B–S1D). We chose the best three assemblies (k-mer 57 of Velvet-Oases, k-mer 61 of SOAPdenovo-trans, k-mer 25 of Trinity) from each software, based on the number of reads used for assembly, N50, average contig length, and the number of 'N' bases. However, unlike in genome assemblies, where the aim is to maximize contig and scaffold lengths, TA's are expected to contain smaller sequence fragments. Hence, to find the best method to produce TAs, we compared the coverage of TAs produced from each assembler to the closest annotated genome, as proposed by *Haas et al. (2013)*. Subsequently 34% and 37% of the TA contigs produced by SOAPdenovoTrans and Velvet-Oases, respectively, but 67**%** of the TA contigs generated by Trinity, aligned to the predicted Protein set of *C. floridanus* with high confidence (100 bp alignment length and 70% sequence identity). There are 3,873 and 5,916 proteins of *C. floridanus,* that are represented by transcripts (TA contigs) with ≥90% alignment coverage, produced by SOAPdenovoTrans and Velvet-Oases, respectively (Fig. 3). Overall, the Trinity-assembled initial-TAs with different k-mers (range: 21–31) covered a higher number (range: 6,704–7,288) of *C. floridanus* proteins than different k-mer-specific Initial-TAs of the other assemblers (SOAPdenovoTrans and Velvet-Oases) with the same criteria. Trinity with k-mer 31, was an exception (5,592 *C. floridanus* protein only); this k-mer performed weakly in general, and only generated 14,661 transcripts (File S1, Tables S1D), i.e., only a 10th of the number generated by the other Trinity k-mers (range: 155,962–189,896 transcripts).

The lower alignment rates of contigs produced by SOAPdenovoTrans and Velvet-Oases, suggest that these methods produce more false positives, and incomplete assemblies than Trinity. In particular, contigs produced by SOAP include strings of N's used for gap closing of read pairs. The number of genes identified by all assemblers (range: 10,916 to 11,123, with a total of 10,144 identified by all three assemblers), and those identified by one or two of them are shown in a Venn diagram (File S1, Fig. S1D). An approach for de novo transcriptome assembly that takes advantage of the assemblies of all different assemblers with various k-mer lengths is highly desirable, but it would also make datasets more complex (high heterozygosity, high error rate, high number of indels, contig orientation), and scaffolding would become computationally intractable. We therefore chose to only use the Trinity assembler as it is able to reconstruct the best full-length transcripts, with the ability to report also splicing variants. Our results agree with other studies that regard Trinity as the best transcriptome assembler over SOAPdenovoTrans and Velvet-Oases (*Zhao et al., 2011*; *Nakasugi et al., 2014*; *He et al., 2015*; *Rana et al., 2016*).

Biological reasons for the misalignment of contigs to *C. floridanus* proteins, including sequences that are highly variable, taxon restricted (orphans), or of non-source species origin (e.g., microbial contaminations, *Johansson et al., 2013*), are expected to affect all assemblers equally and not produce biases in the comparison. Trinity performed the best

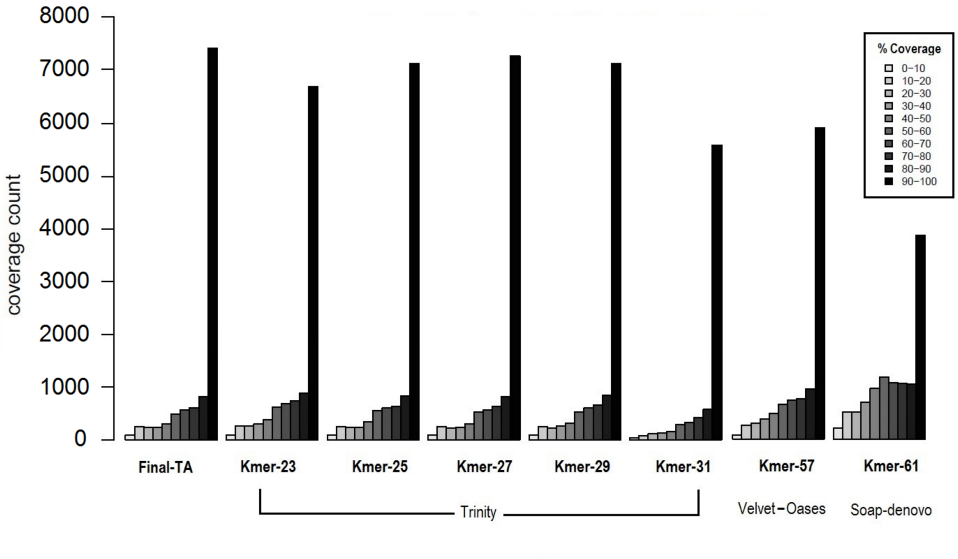

Figure 3 Coverage of contigs to *C. floridanus* predicted proteins. All k-mer assemblies produced by Trinity, and used to merge the Final-TA are shown, whereas only the assembly k-mers producing the largest fraction of >90% coverage are shown for Velvet-Oasis and SOAP-denovo. With respect to the transcripts that covered >90% of the *C. floridanus* proteins, nearly all Trinity k-mer assemblies outperformed Velvet-Oasis and SOAP-denovo, and the Final-TA yielded more transcripts than any individual k-mer assembly.

Full-size DOI: 10.7717/peerj.3998/fig-3

for our data, and produced 7,145 full length transcripts with more than 90% coverage with CflorPP proteins (File S1, Table S1E), consistent with its performance in previous comparisons of TA methods (*Li et al., 2014*; *Chang, Wang & Li, 2014*; *Kiuchi et al., 2014*; *Balakrishnan et al., 2014*; *Davies et al., 2014*). Given that Velvet-Oases has been found to perform nearly as well as Trinity in other studies (*Tulin et al., 2013*), it is likely that different assemblers and settings are optimal for different types of libraries, platforms, and read lengths. Initial-TA contigs from Trinity were chosen to build a combined assembly (Meta-TA) from different k-mer runs (File S1, Table S1F) by scaffolding overlapping contigs with Vmatch software (*Abouelhoda, Kurtz & Ohlebusch, 2002*). This procedure results in a non-redundant set of contig sequences originating from the initial k-mer specific Trinity assemblies. The merged Meta-TA assembly produced a total of 234,970 contigs (File S1, Tables S1G, S1H), that covered nearly 7,435 *C. floridanus* genes. This covers more than 90% of their protein length, which is 2–25% more than the length obtained from any single k-mer run of Trinity, including the default k-mer setting (Fig. 3). Thus, combining k-mer runs enhanced the average length, and length distribution of the contigs (File S1 Tables S1G, S1H).

We achieved the highest-quality transcriptome (Final-TA) by combining the output from the Trinity *de novo* assembler with different k-mers (range: 21–31). We also show that simultaneous assessment of a variety of metrics, not just focusing on the contig length or N50 value, is necessary to gauge the quality of the assemblies (File S1, Tables S1F and S1H, Fig. S1I). For transcriptome assembly, functional evaluation is more important than

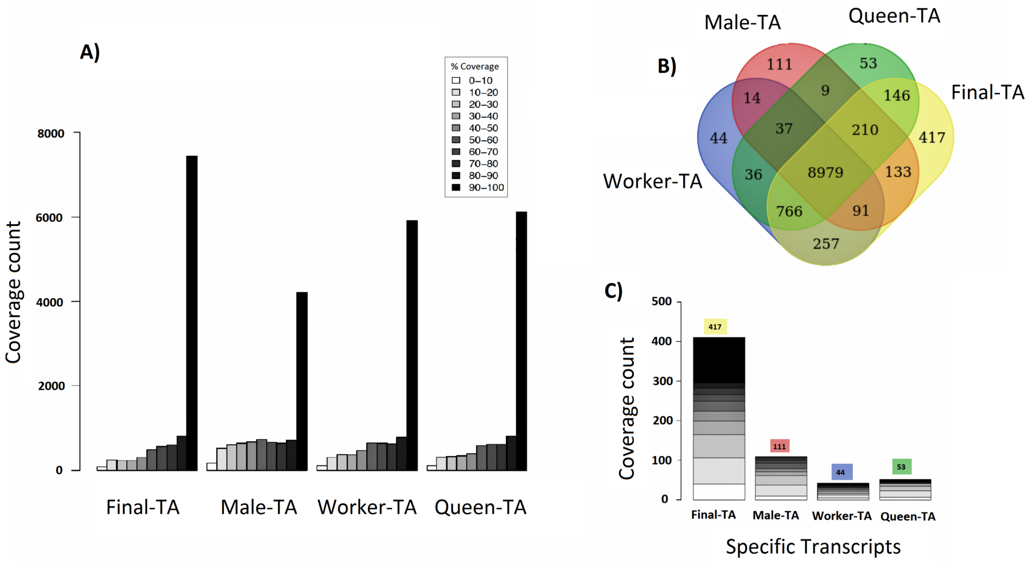

Figure 4 Coverage of contigs between castes. (A) Alignment of caste-specific assemblies to *C. floridanus* predicted proteins compared to the Final-TA, including all castes. (B) Overlap of transcripts aligning to *C. floridanus* proteins, by castes and the Final-TA. (C) Alignment coverage of contigs to *C. floridanus* assembled by caste or in the Final-TA.

Full-size DOI: 10.7717/peerj.3998/fig-4

other parameters, like transcripts length or N50 values (Vera et al., 2008; O'Neil & Emrich, 2013). In addition, for polyphenic species like ants, it is important to sample individuals from all morphs, and from different developmental stages in order to obtain an adequate overview of expressed genes. For example, an assembly constructed with combined data (pooling males, queens and workers) covered nearly 7,435 *C. floridanus* genes (with 90% of their protein length), whereas an assembly of males, queens and workers, separately, each covered fewer genes with 90% coverage cut-off (4,215, 6,134 and 5,913 genes, respectively) (Fig. 4A). Also, a BUSCO transcriptome quality analysis using the Hymenoptera reference lineage showed that the Final-TA produced a higher number (3,936 genes) of full length transcripts than the specific caste assemblies (File S1 Fig. S1J).

The combined assembly (Final-TA) represents the maximum number of full length transcripts, which covers 10,999 *C. floridanus* genes from a total of 15,977 non-redundant proteins (CflorPP90), (File S1, Table S1H). Although the Final-TA covers only 69% of the *C. floridanus* genes, most transcripts share gene orthologies with other ant species, such as *A. echinatior*, *S. invicta*, *A. cephalotes*, *H. saltator*, *L. humile* and *P. parbatus* (Wurm et al., 2009). In total, the 42, 476 contigs in the Final-TA were annotated with 17,496 unique genes across all the 31 ant species available in the NCBI database, and 8,906 proteins top matches came from *C. floridanus* (File S1, Table S1J). In our final annotation the number of unique genes is within the range that we see in other sequenced ant genomes (range: 16,123–18,564; median: 17,220; Oxley et al., 2014). Out of 42,476 contigs, 29,041 share gene orthologies with seven ant genomes (*C. floridanus, A. echinatior*, *S. invicta*, *A. cephalotes*, *H. saltator*, *L. humile* and *P. parbatus*) sequences, and represent 8,375 conserved unique genes across them. The remaining 13,435 contigs showed orthologs unique to one ant species only and were absent in the other ant species.

Caste-specific expressed transcripts

Separate assemblies of the castes showed that 8,979 *C. floridanus* genes (3,725 of which had 90–100% coverage, Fig. 4A), are shared across the castes (Fig. 4B), whereas 536 genes were unique to a caste or sex (133 in males, 257 in workers, 146 in queens), and 766 specific to the two female castes combined (Fig. 4C, Files S3A–S3D). A comparison of assemblies from different castes (male, queen, worker) reveals that the combined assembly has 417 caste-specific gene transcripts (Fig. 4C). We would have missed these gene transcripts if separate assemblies for each sample had been made, given differences in expression levels, and sequencing coverage for each specific caste. Using the Final-TA, we determined caste and sex -specific genes, and found that 766 genes were only expressed in queens and workers (File S3D), whereas 133 of the genes were exclusively expressed in males. Among the female-specific genes, several are involved in biosynthetic processes, such as macromolecule, and lipid biosynthesis, or biopolymer catabolic process (File S1, Fig. S1K). In addition, many are associated with cell communication and signal transduction. The queen-specific genes comprise categories, such as translation, nucleus, carbohydrate binding, nucleic acid binding, system process, ATP-, DNA- and RNA binding. Of the 133 male-specific genes, 42 matched those reported previously in the honey bee sperm proteome (*Zareie et al., 2013*). This subset contains a significant number of proteins that are predicted to act in enzyme regulation or in nucleic acid binding and processing (*Zareie et al., 2013*).

Transcriptome functional annotation & analysis

From the Meta-TA, we selected contigs at least 200 bp in length for annotation, and gene ontology assessment in the Evidence-TA. This resulted in 120,212 contigs in total, including predicted isoforms (Genbank accession number: GFLV00000000). Of these, 39,262 (32.7%) contigs were annotated with available ant genomes including the two closest relatives *C. floridanus* and *L. niger*, and 5010 (4.2%) to other insects (File S4, NCBI update 2015). These comprise the Evidence-TA (Fig. 1). The top hit species distribution from a BLAST analysis, including the NR database update 2017, shows that most annotations come from *C. floridanus* (20,295 transcripts), *L. niger* (7,314 transcripts), *C. biroi* (1,305 transcripts), *L. humile* (1,227 transcripts) and *S. invicta* (1,136 transcripts) (File S1, Fig. S1H).

Of the sequences excluded from the Evidence-TA, 1,847 (1.5% of the contigs in Meta-TA) were aligned to micro-organisms, such as bacteria, viruses, and fungi, or other non-species genes (File S3 and *Johansson et al., 2013*). However, the majority (61.6%) of the contigs (74093) were not aligned to any available databases including ants, other eukaryotes, or their pathogens. In comparison, our annotation rate is within the range reported (20–40%) for several other *de novo* transcriptome assemblies in non-model species (*Vera et al., 2008*; *Wang et al., 2010*; *Fraser et al., 2011*; *Hou et al., 2011*; *Du et al., 2012*), and considerably higher than the rate of unique genes annotated in an earlier study of *F. exsecta* (*Badouin et al., 2013*). Most of the unannotated contigs are shorter (mean 486.7 bp; median 343 bp) than those annotated (mean 1,965 bp and median 1,510 bp, respectively). Only 29% of the unannotated contigs were longer than 500bp. The total size of the annotated contigs is 87 MB base pairs, whereas the unannotated ones only comprise 37 MB base pairs, despite being more numerous. This suggests that many of the unannotated sequences are fragmented.

Our chosen homology-based annotation strategy leads to an inherent bias so that only conserved genes, with identifiable orthologs, are included in further analyses. A subset of these contigs may represent sequences that have gone through high diversification since the divergence between *Formica* and *Camponotus* (*Moreau et al., 2006*), or other ant genes, not preserved in the phylogenetic lineage leading to *Camponotus.* Given the average rate of orphan genes found in arthropod genomes (*Wissler et al., 2013*), and the estimation that every eukaryotic genome contains 10–20% of taxonomically-restricted genes (TRGs) without any significant sequence similarity to genes of other species (*Johnson & Tsutsui, 2011*), our rough estimate is that approximately 1,000 true genes were lost when we filtered out the unannotated contigs. These should comprise a minority of the unannotated transcripts, whereas the majority should be expressed pseudogenes (*Kalyana-Sundaram et al., 2012*), previously uncharacterized transposons or other mobile genetic elements (*Dion-Cote et al., 2014*), micro-organisms such as viruses *Johansson et al., 2013*, regulatory RNA species of protein-coding genes such as micro-RNA (miRNA) precursors and regulatory long non-coding RNA (lncRNA) (*Morris & Mattick, 2014*), and/or prediction artefacts. The predominantly short lengths of the unannotated contigs support this interpretation. A genome sequence of *F. exsecta,* or close relatives, would be necessary for investigating whether the unannotated sequences are indeed protein-coding genes, prediction artifacts, or non-coding RNA. A caste specific role of TRGs has been suggested (*Sumner, Pereboom & Jordan, 2006*; *Ferreira et al., 2013*; *Feldmeyer, Elsner & Foitzik, 2014*; *Harpur et al., 2014*), but is beyond the scope of this study.

When our contigs were aligned to the *C. floridanus* genome, and the Unigene-TA constructed, our data contained 6,894 annotated unique genes, with 3,807 genes in singlets (File S5), and 3,087 genes distributed across several isoforms (range 2–84, average = 4.1, s.d. = 4.2, File S6; Fig. 5). In comparison, in *Drosophila melanogaster* approximately 60% of the multi-exon genes were estimated to contain several isoforms (*Stolc et al., 2004*). As the majority of genes (65%) with many isoforms encodes two or more protein products, alternative splicing generates considerable protein diversity in *Drosophila* (*Misra et al., 2002*). Our data suggests that the ca. 8,000 unique hits to other insect genomes found an earlier transcriptome of the *F. exsecta* (*Badouin et al., 2013*) is an overestimate due to isoforms treated as separate genes. This demonstrates that the steps from Evidence-TA contigs (containing isoforms separately) to Genewise-TA (isoforms combined to genes) are important for the correct estimation of gene numbers.

Considering isoforms as separate genes could potentially bias downstream analyses such as GO-enrichment. A total of 23,016 sequences from the Final-TA were assigned to one or more gene ontology using Blast2Go. Of those sequences, 11,025 belonged to Unigene-TA along with isoforms and 4,560 to Unigene-TA (File S1, Fig. S11). The percentages of gene counts for each gene ontology term were significantly different between the Unigene-TA

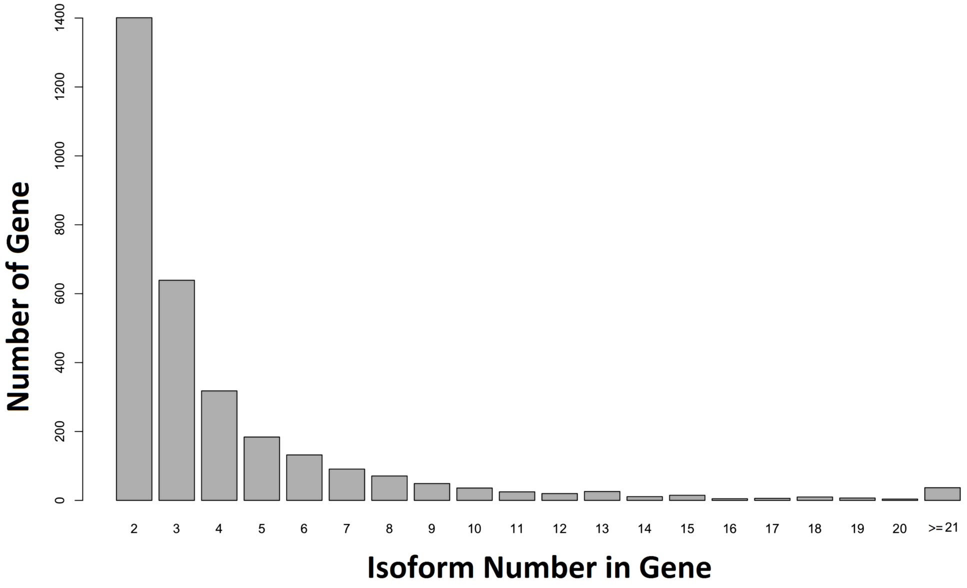

Figure 5 **Number of isoforms detected in non-singlet genes.** More than one isoform was found in 44.71% ($n = 3,087$) of the annotated genes against *C. floridanus*.
Full-size DOI: 10.7717/peerj.3998/fig-5

sets with and without isoforms ($t = 3.4$, $df = 11$, $p = 0.006$) in the cellular component GO category. However, no differences were found in the biological process ($t = -0.95$, $df = 16$, $p = 0.36$), and molecular function ($t = 0.088$, $df = 9$, $p = 0.93$) GO categories. As expected, the frequencies of GO terms obtained from the Final-TA (i.e., the whole data set containing all isoforms) closely follow those found in an earlier paper on the same species (*Badouin et al., 2013*). In both studies the most frequent 12 GO terms are the same within the category biological processes. In addition, the two most highly represented GO terms for molecular function coincide with those detected by *Badouin et al. (2013)*.

Gene families and analysis

The evolution of many gene families is thought to be affected by the evolution of eusociality (*Simola et al., 2013*). In particular, social insects make intensive use of chemical cues in communication and recognition (*Bradbury & Vehrencamp, 2011*; *Richard & Hunt, 2013*; *Wyatt, 2013*), and have therefore been predicted to have diversified in this respect, compared to solitary insects (*Bradbury & Vehrencamp, 2011*; *Wyatt, 2013*). We detected 12 chemosensory protein (CSP) genes, and three odorant binding protein (OBP) genes, in which the number of isoforms varies between 1 and 10 (average = 2.2, s.d. = 2.0, File S7). This falls within the range found in other species of ants, in which the number of functional CSP genes ranges from 11 to 21 depending on species (*Vieira & Rozas, 2011*; *Zhou et al., 2013*; *Kulmuni & Havukainen, 2013*). This is, however surprisingly low, given that solitary insects, such as the flour beetle *Tribolium castaneum* (19 genes), the silkworm *Bombyx mori* (22 genes), and the locust *Locusta migratoria manilensis* (more than 70 genes), show gene numbers in the upper range or much higher than those found in social insects. Thus, the evolution of gene families, and the genetic underpinning of odour cue diversity in social insects, is likely more complex, and should be considered together with isoform variation across taxa. Indeed, presence of multiple isoforms has been demonstrated for several CSP genes. For example, seven CSP isoforms were found in individual *Schistocerca gregaria* legs (*Angeli et al., 1999*), 14 CSP isoforms in *Locusta migratoria* (*Ban et al., 2003*), seven OBP isoforms in the hemolymph of *Tenebrio molitor* (*Graham et al., 2003*) larvae, and 38 isoforms in the fly *Drosophila melanogaster* (*Graham & Davies, 2002*).

The number of immune genes in ants is predicted to depend on the extent of social immunity and hygiene behaviours (*Cremer, Armitage & Schmid-Hempel, 2007*), which may reduce the need for a high number of immune genes (*Evans et al., 2006*). Indeed, in the honey bee, many immune-gene families appear to be depauperate, when compared to solitary insects (*Evans et al., 2006*; *Werren et al., 2010*). Our transcriptome contained only some of the antimicrobial-peptide coding genes described previously in ant genomes (77 unique genes, up to 10 isoforms, average = 1.4, s.d. = 1.4, File S8). This may partly be explained by the fact that none of the samples were purposely challenged with pathogens before preparation. However, given the variety of pathogens (pathogen pressure) present in natural colonies (*Baird, Woolfolk & Watson, 2007*; *Boots et al., 2012*; *Duff et al., 2016*), these genes might not be expressed at the specific time point samples were collected, but may be present in the genome. In general, any conclusions about absence of genes from the genome, and gene family sizes in general should be treated with caution when based on expression data only.

Vitellogenins (Vg) comprise a third gene family (*Wahli et al., 1981*; *Sappington, 2002*; *Tufail & Takeda, 2008*), which has diversified among many ants, and has many genes with caste biased expression patterns (*Wurm, Wang & Keller, 2010*; *Corona et al., 2013*; *Morandin et al., 2014*; *Salmela & Sundström, 2017*). Here we report five copies of Vg and one Vg receptor (File S9), and their isoforms Vitellogenin-like-B (four isoforms), Vitellogenin-like-D (two isoforms), Vitellogenin receptor (four), and only one isoform each of Vitellogenin-like-A, Vitellogenin-like-C, and the conventional Vitellogenin genes. None of these were specific to any sex or caste. Earlier work on *F. exsecta,* where the same reads of queens and workers were applied, found altogether four vitellogenin gene orthologues, (Vitellogenin, Vitellogenin-like-A, Vitellogenin-like-B, Vitellogenin-like-C) each with specific expression patterns and great structural variation (*Morandin et al., 2014*). Vitellogenin-like-D gene is shortest among all Vg's and missed in earlier work due to unavailability of annotation in initial stage and its partial assembly. As vitellogenins are known to have both antioxidant, antimicrobial and storage functions (*Havukainen, Halskau & Amdam, 2011*), in addition to their key role in caste regulation in e.g., the honeybee (*Piulachs et al., 2003*; *Havukainen, Halskau & Amdam, 2011*; *Salmela & Sundström, 2017*), further work on isoform expression could shed light on the role of the different vitellogenins in the of caste phenotypes (*Wurm, Wang & Keller, 2010*; *Corona et al., 2013*; *Morandin et al., 2014*).

Genes related to epigenetic regulation are of current interest for their role in generating and maintaining caste differences in social insects. Genomic analyses have revealed the evolutionary persistence of DNA methylation across Hymenoptera (*Glastad et al., 2011*), in contrast to e.g., Diptera, where methylation is diminished (*Marhold et al., 2004*). Our transcriptome data reveals all DNA methyltransferases required for CpG methylation, including DNMT1 for maintaining methylation patterns, and DNMT3, which is responsible for *de novo* methylation. These DNMTs have also been found in other ant species (*Bonasio et al., 2010*; *Suen et al., 2011*; *Smith et al., 2011b*; *Wurm et al., 2011*; *Nygaard et al., 2011*; *Oxley et al., 2014*; *Schrader et al., 2014*). DNMT2 RNA methyltransferase was recently reported to be required for the establishment and hereditary maintenance of methylation patterns in mice (*Kiani et al., 2013*). Thus, the epigenetic inheritance and its regulation is undoubtedly more complex than understood today (see also e.g., *Wang et al., 2013*; *Yan et al., 2014*). Similarly to e.g., mice (*Lin et al., 2000*), we found splice variants in DNMT genes (up to 4 isoforms cf. File S7), and given the role of methylation in caste differentiation in social insects (*Glastad et al., 2011*), the role of splice variants is certainly an interesting direction for more studies.

CONCLUSIONS

Given the rapid advances in transcriptomics of non-model organisms, with ever more comparative studies being carried out, it is all the more important that the transcriptomes are reliable. In this study we provide insights into essential criteria that should be taken into account for a reliable transcriptome assembly without a reference genome. First, we find that analyses relying on single life stages or morphs is likely to miss some genes due to low sequencing depth or expression at specific stage. Second, care should be taken in choosing the assembly methods, and different methods should be compared in order to find which one performs the best for a given species data set. In addition to this, filtering with a reference genome of a related species is necessary for quantifying transcript abundance, and doing functional and structural annotation of transcripts, so that isoforms are not analyzed as unique genes in downstream applications. This necessarily comes with the cost of losing orphan genes from the data. While the study of orphan genes is obviously valuable, for the purposes of facilitating species comparisons on various taxonomical scales, we opted for analyzing genes where annotations are available so that we can be sure we are working with true genes. Identifying isoform variation from unique genes allows studies on structural mRNA diversity in genes of interest, which may be evolutionarily more important than variation in expression levels.

ACKNOWLEDGEMENTS

We thank Pekka Pamilo, Helena Johansson and Abhilash Nair for highly useful comments on the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded by the Academy of Finland (to Centre of Excellence in Biological Interactions, grants no. 252411 and 284666, to Heikki Helanterä 135970, 140990, to Kalevi

Trontti 134561, and Christopher Wheat 131155). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Academy of Finland: 252411, 284666, 135970, 140990, 134561, 131155.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Kishor Dhaygude and Kalevi Trontti conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Jenni Paviala and Claire Morandin contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Christopher Wheat and Liselotte Sundström conceived and designed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Heikki Helanterä conceived and designed the experiments, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

BioProject: PRJNA213662
BioSample: SAMN03799245
Figshare: https://doi.org/10.6084/m9.figshare.4877360.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3998#supplemental-information.

REFERENCES

Abouelhoda MI, Kurtz S, Ohlebusch E. 2002. The enhanced suffix array and its applications to genome analysis. In: Guigó R, Gusfield D, eds. *Algorithms in bioinformatics. WABI 2002. Lecture notes in computer science*, vol. 2452. Berlin, Heidelberg: Springer.

Abouheif E. 2002. Evolution of the gene network underlying wing polyphenism in ants. *Science* **297**:249–252 DOI 10.1126/science.1071468.

Alvarado S, Rajakumar R, Abouheif E, Szyf M. 2015. Epigenetic variation in the Egfr gene generates quantitative variation in a complex trait in ants. *Nature Communications* **6**:6513 DOI 10.1038/ncomms7513.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. *Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc*.

Angeli S, Ceron F, Scaloni A, Monti M, Monteforti G, Minnocci A, Petacchi R, Pelosi P. 1999. Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria*. *European Journal of Biochemistry* **262**:745–754 DOI 10.1046/j.1432-1327.1999.00438.x.

Ashby R, Forêt S, Searle I, Maleszka R. 2016. MicroRNAs in honey bee caste determination. *Scientific Reports* **6**:18794 DOI 10.1038/srep18794.

Badouin H, Belkhir K, Gregson E, Galindo J, Sundström L, Martin SSJ, Butlin RKR, Smadja CM. 2013. Transcriptome characterisation of the ant *Formica exsecta* with new insights into the evolution of desaturase genes in social hymenoptera. **8**:e68200 DOI 10.1371/journal.pone.0068200.

Baird R, Woolfolk S, Watson CE. 2007. Survey of bacterial and fungal associates of black/hybrid imported fire ants from mounds in Mississippi. *Southeastern Naturalist* **6**:615–632 DOI 10.1656/1528-7092(2007)6[615:SOBAFA]2.0.CO;2.

Balakrishnan CN, Mukai M, Gonser RA, Wingfield JC, London SE, Tuttle EM, Clayton DF. 2014. Brain transcriptome sequencing and assembly of three songbird model systems for the study of social behavior. *PeerJ* **2**:e396 DOI 10.7717/peerj.396.

Ban L, Scaloni A, Brandazza A, Angeli S, Zhang L, Yan Y, Pelosi P. 2003. Chemosensory proteins of *Locusta migratoria*. *Insect Molecular Biology* **12**:125–134 DOI 10.1046/j.1365-2583.2003.00394.x.

Berens AJ, Hunt JH, Toth AL. 2015. Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. *Molecular Biology and Evolution* **32**:690–703 DOI 10.1093/molbev/msu330.

Bonasio R. 2014. The role of chromatin and epigenetics in the polyphenisms of ant castes. *Briefings in Functional Genomics* **13**:235–245 DOI 10.1093/bfgp/elt056.

Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, Zhang P, Huang Z, Berger SL, Reinberg D, Wang J, Liebig J. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329**:1068–1071 DOI 10.1126/science.1192428.

Bonfert T, Csaba G, Zimmer R, Friedel CC. 2013. Mining RNA–Seq data for infections and contaminations. *PLOS ONE* **8**:e73071 DOI 10.1371/journal.pone.0073071.

Boots B, Keith AM, Niechoj R, Breen J, Schmidt O, Clipson N. 2012. Unique soil microbial assemblages associated with grassland ant species with different nesting and foraging strategies. *Pedobiologia* **55**:33–40 DOI 10.1016/j.pedobi.2011.10.004.

Bradbury J, Vehrencamp J. 2011. *Principles of animal communication*. Sunderland: Sinauer Associates.

Brown WD, Keller L, Sundström L. 2002. Sex allocation in mound-building ants: the roles of resources and queen replenishment. *Ecology* **83**:1945–1952 DOI 10.1890/0012-9658(2002)083[1945:SAIMBA]2.0.CO;2.

Brown W, Liautard C, Keller L. 2003. Sex-ratio dependent execution of queens in polygynous colonies of the ant *Formica exsecta*. *Oecologia* **134**:12–17 DOI 10.1007/s00442-002-1072-8.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**:25–36 DOI 10.1016/j.cell.2008.06.030.

Chang Z, Wang Z, Li G. 2014. The impacts of read length and transcriptome complexity for de novo assembly: a simulation study. *PLOS ONE* **9**:e94825 DOI 10.1371/journal.pone.0094825.

Colgan TJ, Carolan JC, Bridgett SJ, Sumner S, Blaxter ML, Brown MJ. 2011. Polyphenism in social insects: insights from a transcriptome-wide analysis of gene expression in the life stages of the key pollinator, *Bombus terrestris*. *BMC Genomics* **12**:623 DOI 10.1186/1471-2164-12-623.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**:3674–3676 DOI 10.1093/bioinformatics/bti610.

Corona M, Libbrecht R, Wurm Y, Riba-Grognuz O, Studer RA, Keller L. 2013. Vitellogenin underwent subfunctionalization to acquire caste and behavioral specific expression in the harvester ant *Pogonomyrmex barbatus*. *PLOS Genetics* **9**:e1003730 DOI 10.1371/journal.pgen.1003730.

Cremer S, Armitage SAO, Schmid-Hempel P. 2007. Social immunity. *Current Biology* **17**:R693–R702 DOI 10.1016/j.cub.2007.06.008.

Davies KTJ, Tsagkogeorga G, Bennett NC, Dávalos LM, Faulkes CG, Rossiter SJ. 2014. Molecular evolution of growth hormone and insulin-like growth factor 1 receptors in long-lived, small-bodied mammals. *Gene* **549**:228–236 DOI 10.1016/j.gene.2014.07.061.

Dion-Cote A-M, Renaut S, Normandeau E, Bernatchez L. 2014. RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. *Molecular Biology and Evolution* **31**:1188–1199 DOI 10.1093/molbev/msu069.

Du H, Bao Z, Hou R, Wang S, Su H, Yan J, Tian M, Li Y, Wei W, Lu W, Hu X, Wang S, Hu J. 2012. Transcriptome sequencing and characterization for the sea cucumber *Apostichopus japonicus* (Selenka, 1867). *PLOS ONE* **7**:e33311 DOI 10.1371/journal.pone.0033311.

Duff LB, Urichuk TM, Hodgins LN, Young JR, Untereiner WA. 2016. Diversity of fungi from the mound nests of *Formica ulkei* and adjacent non-nest soils. *Canadian Journal of Microbiology* **62**:562–571 DOI 10.1139/cjm-2015-0628.

Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN, Hagen DE. 2016. Hymenoptera genome database: integrating genome annotations in HymenopteraMine. *Nucleic Acids Research* **44**:D793–D800 DOI 10.1093/nar/gkv1208.

Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, De Graaf DC, Debyser G, Deng J, Devreese B, Elhaik E, Evans JD, Foster LJ, Graur D, Guigo R, Hoff K, Holder ME, Hudson ME, Hunt GJ, Jiang H, Joshi V, Khetani RS, Kosarev P, Kovar CL, Ma J, Maleszka R, Moritz RFA, Munoz-Torres MC, Murphy TD, Muzny DM, Newsham IF, Reese JT, Robertson HM, Robinson GE, Rueppell O, Solovyev V, Stanke M, Stolle E, Tsuruda JM, Vaerenbergh M, Waterhouse RM, Weaver DB, Whitfield CW, Wu Y, Zdobnov EM, Zhang L, Zhu D, Gibbs RA. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* **15**86 DOI 10.1186/1471-2164-15-86.

Evans JD, Aronstein K, Chen YP, Hetru C, Imler J-L, Jiang H, Kanost M, Thompson GJ, Zou Z, Hultmark D. 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Molecular Biology* **15**:645–656 DOI 10.1111/j.1365-2583.2006.00682.x.

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**:151–161 DOI 10.1111/mec.12490.

Ferreira PG, Patalano S, Chauhan R, Ffrench-Constant R, Gabaldón T, Guigó R, Sumner S. 2013. Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. *Genome Biology* **14**:R20 DOI 10.1186/gb-2013-14-2-r20.

Fournier D, Estoup A, Orivel J, Foucaud J, Jourdan H, Le BretonJ, Keller L. 2005. Clonal reproduction by males and females in the little fire ant. *Nature* **435**:1230–1234 DOI 10.1038/nature03705.

Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA. 2011. Sequencing and characterization of the guppy (Poecilia reticulata) transcriptome. *BMC Genomics* **12**:202 DOI 10.1186/1471-2164-12-202.

Gadau J, Helmkampf M, Nygaard S, Roux J, Simola DF, Smith CR, Suen G, Wurm Y, Smith CD. 2012. The genomic impact of 100 million years of social evolution in seven ant species. *Trends in Genetics* **28**:14–21 DOI 10.1016/j.tig.2011.08.005.

Glastad KM, Hunt BG, Yi SV, Goodisman MAD. 2011. DNA methylation in insects: on the brink of the epigenomic era. *Insect Molecular Biology* **20**:553–565 DOI 10.1111/j.1365-2583.2011.01092.x.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**:644–652 DOI 10.1038/nbt.1883.

Gräff J, Jemielity S, Parker JD, Parker KM, Keller L. 2007. Differential gene expression between adult queens and workers in the ant *Lasius niger*. *Molecular Ecology* **16**:675–683 DOI 10.1111/j.1365-294X.2007.03162.x.

Graham LA, Brewer D, Lajoie G, Davies PL. 2003. Characterization of a subfamily of beetle odorant-binding proteins found in hemolymph. *Molecular & Cellular Proteomics* **2**:541–549 DOI 10.1074/mcp.M300018-MCP200.

Graham LA, Davies PL. 2002. The odorant-binding proteins of *Drosophila melanogaster*: annotation and characterization of a divergent gene family. *Gene* **292**:43–55 DOI 10.1016/S0378-1119(02)00672-8.

Gramates LS, Marygold SJ, Santos G dos, Urbano J-M, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB, Falls K, Goodman JL, Hu Y, Ponting L, Schroeder AJ, Strelets VB, Thurmond J, Zhou P, FlyBase Consortium. 2017. FlyBase at 25: looking to the future. *Nucleic Acids Research* **45**:D663–D671 DOI 10.1093/nar/gkw1016.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, Van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin

W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE. 2011. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**:473–479 DOI 10.1038/nature09715.

Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**:965–978 DOI 10.1016/j.infsof.2005.09.005.

Haag-Liautard C, Vitikainen E, Keller L, Sundström L. 2009. Fitness and the level of homozygosity in a social insect. *Journal of Evolutionary Biology* **22**:134–142 DOI 10.1111/j.1420-9101.2008.01635.x.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**:1494–1512 DOI 10.1038/nprot.2013.084.

Hannon. 2010. FASTX toolkit. *Available at http://hannonlab.cshl.edu/fastx_toolkit/index.html*.

Harpur BA, Kent CF, Molodtsova D, Lebon JMD, Alqarni AS, Owayss AA, Zayed A. 2014. Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences of the United States of America* **111**:2614–2619 DOI 10.1073/pnas.1315506111.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, Van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. 2012. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Research* **22**:1760–1774 DOI 10.1101/gr.135350.111.

Havukainen H, Halskau Ø, Amdam GV. 2011. Social pleiotropy and the molecular evolution of honey bee vitellogenin. *Molecular Ecology* **20**:5111–5113 DOI 10.1111/j.1365-294X.2011.05351.x.

He B, Zhao S, Chen Y, Cao Q, Wei C, Cheng X, Zhang Y. 2015. Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. *BMC Genomics* **16**:65 DOI 10.1186/s12864-014-1192-7.

Helms Cahan S, Keller L. 2003. Complex hybrid origin of genetic caste determination in harvester ants. *Nature* **424**:306–309 DOI 10.1038/nature01744.

Hornett EA, Wheat CW. 2012. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC Genomics* **13**:361 DOI 10.1186/1471-2164-13-361.

Hou R, Bao Z, Wang S, Su H, Li Y, Du H, Hu J, Wang S, Hu X. 2011. Transcriptome sequencing and de novo analysis for yesso scallop (*Patinopecten yessoensis*) using 454 GS FLX. *PLOS ONE* **6**:e21560 DOI 10.1371/journal.pone.0021560.

Hunt BG, Goodisman MAD. 2010. Evolutionary variation in gene expression is associated with dimorphism in eusocial vespid wasps. *Insect Molecular Biology* **19**:641–652 DOI 10.1111/j.1365-2583.2010.01021.x.

Hunt BG, Ometto L, Wurm Y, Shoemaker D, Yi SV, Keller L, Goodisman MAD. 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**:15936–15941 DOI 10.1073/pnas.1104825108.

Hunt BG, Wyder S, Elango N, Werren JH, Zdobnov EM, Yi SV, Goodisman MAD. 2010. Sociality is linked to rates of protein evolution in a highly social insect. *Molecular Biology and Evolution* **27**:497–500 DOI 10.1093/molbev/msp225.

Johansson H, Dhaygude K, Lindström S, Helanterä H, Sundström L, Trontti K. 2013. A metatranscriptomic approach to the identification of microbiota associated with the ant *Formica exsecta*. *PLOS ONE* **8**:e79777 DOI 10.1371/journal.pone.0079777.

Johnson BR, Tsutsui ND. 2011. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC Genomics* **12**:164 DOI 10.1186/1471-2164-12-164.

Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu Y-M, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ, Iyer MK, Barrette T, Shanmugam A, Dhanasekaran SM, Palanisamy N, Chinnaiyan AM. 2012. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **149**:1622–1634 DOI 10.1016/j.cell.2012.04.041.

Kapheim KM, Pan H, Li C, Salzberg SL, Puiu D, Magoc T, Robertson HM, Hudson ME, Venkat A, Fischman BJ, Hernandez A, Yandell M, Ence D, Holt C, Yocum GD, Kemp WP, Bosch J, Waterhouse RM, Zdobnov EM, Stolle E, Kraus FB, Helbing S, Moritz RFA, Glastad KM, Hunt BG, Goodisman MAD, Hauser F, Grimmelikhuijzen CJP, Pinheiro DG, Nunes FMF, Soares MPM, Tanaka ÉD, Simões ZLP, Hartfelder K, Evans JD, Barribeau SM, Johnson RM, Massey JH, Southey BR, Hasselmann M, Hamacher D, Biewer M, Kent CF, Zayed A, Blatti C, Sinha S, Johnston JS, Hanrahan SJ, Kocher SD, Wang J, Robinson GE, Zhang G. 2015. Genomic signatures of evolutionary transitions from solitary to group living. *Science* **348**:1139–1143 DOI 10.1126/science.aaa4788.

Kiani J, Grandjean V, Liebers R, Tuorto F, Ghanbarian H, Lyko F, Cuzin F, Rassoulzadegan M. 2013. RNA–mediated epigenetic heredity requires the cytosine methyltransferase Dnmt2. *PLOS Genetics* **9**:e1003498 DOI 10.1371/journal.pgen.1003498.

Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, Ishihara G, Kawaoka S, Sugano S, Shimada T, Suzuki Y, Suzuki MG, Katsuma S. 2014. A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* **509**:633–636 DOI 10.1038/nature13315.

Kocher SD, Li C, Yang W, Tan H, Yi SV, Yang X, Hoekstra HE, Zhang G, Pierce NE, Yu DW. 2013. The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. *Genome Biology* **14**:R142 DOI 10.1186/gb-2013-14-12-r142.

Konorov EA, Nikitin MA, Mikhailov KV, Lysenkov SN, Belenky M, Chang PL, Nuzhdin SV, Scobeyeva VA. 2017. Genomic exaptation enables *Lasius niger* adaptation to urban environments. *BMC Evolutionary Biology* **17**:39 DOI 10.1186/s12862-016-0867-x.

Kulmuni J, Havukainen H. 2013. Insights into the evolution of the CSP gene family through the integration of evolutionary analysis and comparative protein modeling. *PLOS ONE* **8**:e63688 DOI 10.1371/journal.pone.0063688.

Kummerli R, Keller L. 2007. Reproductive specialization in multiple-queen colonies of the ant *Formica exsecta*. *Behavioral Ecology* **18**:375–383 DOI 10.1093/beheco/arl088.

Lattorff HMG, Moritz RFA. 2013. Genetic underpinnings of division of labor in the honeybee (*Apis mellifera*). *Trends in Genetics* **29**:641–648 DOI 10.1016/j.tig.2013.08.002.

Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, Dewey CN. 2014. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology* **15**:553 DOI 10.1186/s13059-014-0553-5.

Lin MJ, Lee TL, Hsu DW, Shen CK. 2000. One-codon alternative splicing of the CpG MTase Dnmt1 transcript in mouse somatic cells. *FEBS Letters* **469**:101–114 DOI 10.1016/S0014-5793(00)01254-0.

Marhold J, Rothe N, Pauli A, Mund C, Kuehle K, Brueckner B, Lyko F. 2004. Conservation of DNA methylation in dipteran insects. *Insect Molecular Biology* **13**:117–123 DOI 10.1111/j.0962-1075.2004.00466.x.

Martin SJ, Helanterä H, Kiss K, Lee YR, Drijfhout FP. 2009. Polygyny reduces rather than increases nestmate discrimination cue diversity in *Formica exsecta* ants. *Insectes Sociaux* **56**:375–383 DOI 10.1007/s00040-009-0035-z.

Martin SJ, Vitikainen E, Helanterä H, Drijfhout FP. 2008. Chemical basis of nest-mate discrimination in the ant *Formica exsecta*. *Proceedings of the Royal Society B: Biological Sciences* **275**:1271–1278 DOI 10.1098/rspb.2007.1708.

Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, Smith CD, Tupy JL, Whitfied EJ, Bayraktaroglu L, Berman BP, Bettencourt BR, Celniker SE, De Grey ADNJ, Drysdale RA, Harris NL, Richter J, Russo S, Schroeder AJ, Shu SQ, Stapleton M, Yamada C, Ashburner M, Gelbart WM, Rubin GM, Lewis SE. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biology* **3**:RESEARCH0083 DOI 10.1186/GB-2002-3-12-RESEARCH0083.

Morandin C, Dhaygude K, Paviala J, Trontti K, Wheat C, Helanterä H. 2015. Caste-biases in gene expression are specific to developmental stage in the ant *Formica exsecta*. *Journal of Evolutionary Biology* **28**:1705–1718 DOI 10.1111/jeb.12691.

Morandin C, Havukainen H, Kulmuni J, Dhaygude K, Trontti K, Helanterä H. 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**:2181–2193 DOI 10.1093/molbev/msu171.

Moreau CS, Bell CD, Vila R, Archibald SB, Pierce NE. 2006. Phylogeny of the ants: diversification in the age of angiosperms. *Science* **312(5770)**:101–104 DOI 10.1126/science.1124891.

Morris KV, Mattick JS. 2014. The rise of regulatory RNA. *Nature Reviews Genetics* **15**:423–437 DOI 10.1038/nrg3722.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. 2011. Sequence-specific error profile of illumina sequencers. *Nucleic Acids Research* **39**:e90–e90 DOI 10.1093/nar/gkr344.

Nakasugi K, Crowhurst R, Bally J, Waterhouse P, Goepfert S. 2014. Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. *PLOS ONE* **9**:e91776 DOI 10.1371/journal.pone.0091776.

Nygaard S, Zhang G, Schiøtt M, Li C, Wurm Y, Hu H, Zhou J, Ji L, Qiu F, Rasmussen M, Pan H, Hauser F, Krogh A, Grimmelikhuijzen CJP, Wang J, Boomsma JJ. 2011. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Research* **21**:1339–1348 DOI 10.1101/gr.121392.111.

Ometto L, Shoemaker D, Ross KG, Keller L. 2011. Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. *Molecular Biology and Evolution* **28**:1381–1392 DOI 10.1093/molbev/msq322.

O'Neil ST, Emrich SJ. 2013. Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC Genomics* **14**:465 DOI 10.1186/1471-2164-14-465.

Oxley PR, Ji L, Fetter-Pruneda I, McKenzie SK, Li C, Hu H, Zhang G, Kronauer DJC. 2014. The genome of the clonal raider ant *Cerapachys biroi*. *Current Biology* **24**:451–458 DOI 10.1016/j.cub.2014.01.018.

Pearcy M, Aron S, Doums C, Keller L. 2004. Conditional use of sex and parthenogenesis for worker and queen production in ants. *Science* **306**:1780–1783 DOI 10.1126/science.1105453.

Piulachs MD, Guidugli KR, Barchuk AR, Cruz J, Simões ZLP, Bellés X. 2003. The vitellogenin of the honey bee, *Apis mellifera*: structural analysis of the cDNA and expression studies. *Insect Biochemistry and Molecular Biology* **33**:459–465 DOI 10.1016/S0965-1748(03)00021-3.

Rana SB, Zadlock FJ, Zhang Z, Murphy WR, Bentivegna CS, Zhang Z, Dewey C. 2016. Comparison of De Novo transcriptome assemblers and k-mer strategies using the killifish, *Fundulus heteroclitus*. *PLOS ONE* **11**:e0153104 DOI 10.1371/journal.pone.0153104.

Richard F-J, Hunt JH. 2013. Intracolony chemical communication in social insects. *Insectes Sociaux* **60**:275–291 DOI 10.1007/s00040-013-0306-6.

**Sadd BM, Barribeau SM, Bloch G, De Graaf DC, Dearden P, Elsik CG, Gadau J, Grimmelikhuijzen CJ, Hasselmann M, Lozier JD, Robertson HM, Smagghe G, Stolle E, Van Vaerenbergh M, Waterhouse RM, Bornberg-Bauer E, Klasberg S, Bennett AK, Câmara F, Guigó R, Hoff K, Mariotti M, Munoz-Torres M, Murphy T, Santesmasses D, Amdam GV, Beckers M, Beye M, Biewer M, Bitondi MM,

Blaxter ML, Bourke AF, Brown MJ, Buechel SD, Cameron R, Cappelle K, Carolan JC, Christiaens O, Ciborowski KL, Clarke DF, Colgan TJ, Collins DH, Cridge AG, Dalmay T, Dreier S, Du Plessis L, Duncan E, Erler S, Evans J, Falcon T, Flores K, Freitas FC, Fuchikawa T, Gempe T, Hartfelder K, Hauser F, Helbing S, Humann FC, Irvine F, Jermiin LS, Johnson CE, Johnson RM, Jones AK, Kadowaki T, Kidner JH, Koch V, Köhler A, Kraus FB, Lattorff HMG, Leask M, Lockett GA, Mallon EB, Antonio DSM, Marxer M, Meeus I, Moritz RF, Nair A, Näpflin K, Nissen I, Niu J, Nunes FM, Oakeshott JG, Osborne A, Otte M, Pinheiro DG, Rossié N, Rueppell O, Santos CG, Schmid-Hempel R, Schmitt BD, Schulte C, Simões ZL, Soares MP, Swevers L, Winnebeck EC, Wolschin F, Yu N, Zdobnov EM, Aqrawi PK, Blankenburg KP, Coyle M, Francisco L, Hernandez AG, Holder M, Hudson ME, Jackson L, Jayaseelan J, Joshi V, Kovar C, Lee SL, Mata R, Mathew T, Newsham IF, Ngo R, Okwuonu G, Pham C, Pu L-L, Saada N, Santibanez J, Simmons D, Thornton R, Venkat A, Walden KK, Wu Y-Q, Debyser G, Devreese B, Asher C, Blommaert J, Chipman AD, Chittka L, Fouks B, Liu J, O'Neill MP, Sumner S, Puiu D, Qu J, Salzberg SL, Scherer SE, Muzny DM, Richards S, Robinson GE, Gibbs RA, Schmid-Hempel P, Worley KC. 2015. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biology* **16**:76 DOI 10.1186/s13059-015-0623-3.

Salmela H, Sundström L. 2017. Vitellogenin in inflammation and immunity in social insects. *Inflammation and Cell Signaling* **4**:e1506 DOI 10.14800/ICS.1506.

Sappington TW. 2002. The major yolk proteins of higher diptera are homologs of a class of minor yolk proteins in lepidoptera. *Journal of Molecular Evolution* **55**:470–475 DOI 10.1007/s00239-002-2342-0.

Schrader L, Kim JW, Ence D, Zimin A, Klein A, Wyschetzki K, Weichselgartner T, Kemena C, Stökl J, Schultner E, Wurm Y, Smith CD, Yandell M, Heinze J, Gadau J, Oettler J. 2014. Transposable element islands facilitate adaptation to novel environments in an invasive species. *Nature Communications* **5**:5495 DOI 10.1038/ncomms6495.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**:1086–1092 DOI 10.1093/bioinformatics/bts094.

Schwander T, Lo N, Beekman M, Oldroyd BP, Keller L. 2010. Nature versus nurture in social insect caste differentiation. *Trends in Ecology & Evolution* **25**:275–282 DOI 10.1016/j.tree.2009.12.001.

Shbailat SJ, Khila A, Abouheif E. 2010. Correlations between spatiotemporal changes in gene expression and apoptosis underlie wing polyphenism in the ant *Pheidole morrisi*. *Evolution & Development* **12**:580–591 DOI 10.1111/j.1525-142X.2010.00443.x.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212 DOI 10.1093/bioinformatics/btv351.

Simola DF, Graham RJ, Brady CM, Enzmann BL, Desplan C, Ray A, Zwiebel LJ, Bonasio R, Reinberg D, Liebig J, Berger SL. 2016. Epigenetic (re)programming of caste-specific behavior in the ant *Camponotus floridanus*. *Science* **351**:aac6633–aac6633 DOI 10.1126/science.aac6633.

Simola DF, Wissler L, Donahue G, Waterhouse RM, Helmkampf M, Roux J, Nygaard S, Glastad KM, Hagen DE, Viljakainen L, Reese JT, Hunt BG, Graur D, Elhaik E, Kriventseva EV, Wen J, Parker BJ, Cash E, Privman E, Childers CP, Munoz-Torres MC, Boomsma JJ, Bornberg-Bauer E, Currie CR, Elsik CG, Suen G, Goodisman MAD, Keller L, Liebig J, Rawls A, Reinberg D, Smith CD, Smith CR, Tsutsui N, Wurm Y, Zdobnov EM, Berger SL, Gadau J. 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Research* **23**:1235–1247 DOI 10.1101/gr.155408.113.

Simpson SJ, Sword GA, Lo N. 2011. Polyphenism in insects. *Current Biology* **21**:R738–R749 DOI 10.1016/j.cub.2011.06.006.

Smith CRD, Smith CRD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, Fave M-J, Fernandes V, Gibson JD, Graur D, Gronenberg W, Grubbs KJ, Hagen DE, Viniegra ASI, Johnson BR, Johnson RM, Khila A, Kim JW, Mathis KA, Munoz-Torres MC, Murphy MC, Mustard JA, Nakamura R, Niehuis O, Nigam S, Overson RP, Placek JE, Rajakumar R, Reese JT, Suen G, Tao S, Torres CW, Tsutsui ND, Viljakainen L, Wolschin F, Gadau J. 2011a. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proceedings of the National Academy of Sciences of the United States of America* **108**:5667–5672 DOI 10.1073/pnas.1007901108.

Smith CRD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, Fave M-J, Fernandes V, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Helmkampf M, Holley J-A, Hu H, Viniegra ASI, Johnson BR, Johnson RM, Khila A, Kim JW, Laird J, Mathis KA, Moeller JA, Munoz-Torres MC, Murphy MC, Nakamura R, Nigam S, Overson RP, Placek JE, Rajakumar R, Reese JT, Robertson HM, Smith CRD, Suarez AV, Suen G, Suhr EL, Tao S, Torres CW, Van Wilgenburg E, Viljakainen L, Walden KKO, Wild AL, Yandell M, Yorke JA, Tsutsui ND. 2011b. 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**:5673–5678 DOI 10.1073/pnas.1008617108.

Stolc V, Gauhar Z, Mason C, Halasz G, Van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, Bussemaker HJ, White KP. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**:655–660 DOI 10.1126/science.1101312.

Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, Denas O, Elhaik E, Favé M-J, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Harkins TT, Helmkampf M, Hu H, Johnson BR, Kim J, Marsh SE, Moeller JA, Muñoz Torres MC, Murphy MC, Naughton MC, Nigam S, Overson R, Rajakumar R, Reese JT, Scott JJ, Smith CR, Tao S,

Tsutsui ND, Viljakainen L, Wissler L, Yandell MD, Zimmer F, Taylor J, Slater SC, Clifton SW, Warren WC, Elsik CG, Smith CD, Weinstock GM, Gerardo NM, Currie CR. 2011. the genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLOS Genetics* **7**:e1002007 DOI 10.1371/journal.pgen.1002007.

Sumner S, Pereboom JJ, Jordan WC. 2006. Differential gene expression and phenotypic plasticity in behavioural castes of the primitively eusocial wasp, *Polistes canadensis*. *Proceedings of the Royal Society B: Biological Sciences* **273**:19–26 DOI 10.1098/rspb.2005.3291.

Sundström L, Chapuisat M, Keller L. 1996. Conditional manipulation of sex ratios by ant workers: a test of kin selection theory. *Science* **274**:993–995 DOI 10.1126/science.274.5289.993.

Sundström L, Keller L, Chapuisat M. 2003. Inbreeding and sex-biased gene flow in the ant *Formica exsecta*. *Evolution; International Journal of Organic Evolution* **57**:1552–1561 DOI 10.1111/j.0014-3820.2003.tb00363.x.

Surget-Groba Y, Montoya-Burgos JI. 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Research* **20**:1432–1440 DOI 10.1101/gr.103846.109.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**:692–702 DOI 10.1038/nrg3053.

The ENCODE Consortium. 2011. Standards, guidelines and best practices for RNA-Seq V1.0. *Available at https://genome.ucsc.edu/encode/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf*.

Tsutsui ND, Suarez AV, Spagna JC, Johnston JS. 2008. The evolution of genome size in ants. *BMC Evolutionary Biology* **8**:64 DOI 10.1186/1471-2148-8-64.

Tufail M, Takeda M. 2008. Molecular characteristics of insect vitellogenins. *Journal of Insect Physiology* **54**:1447–1458 DOI 10.1016/j.jinsphys.2008.08.007.

Tulin S, Aguiar D, Istrail S, Smith J. 2013. A quantitative reference transcriptome for Nematostella vectensis early embryonic development: a pipeline for de novo assembly in emerging model systems. *EvoDevo* **4**:16 DOI 10.1186/2041-9139-4-16.

Van Zweden JS, Vitikainen E, D'Ettorre P, Sundström L. 2011. Do cuticular hydrocarbons provide sufficient information for optimal sex allocation in the ant *Formica exsecta*? *Journal of Chemical Ecology* **37**:1365–1373 DOI 10.1007/s10886-011-0038-x.

Vera JC, Wheat C, Fescemyer H, Frilander MJ, Crawford DL, Hanski I, Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* **17**:1636–1647 DOI 10.1111/j.1365-294X.2008.03666.x.

Vieira FG, Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biology and Evolution* **3**:476–490 DOI 10.1093/gbe/evr033.

Vitikainen E, Haag-Liautard C, Sundström L. 2011. Inbreeding and reproductive investment in the ant *Formica exsecta*. *Evolution* **65**:2026–2037 DOI 10.1111/j.1558-5646.2011.01273.x.

Vitikainen EIK, Haag-Liautard C, Sundström L. 2015. Natal dispersal, mating patterns, and inbreeding in the ant *Formica exsecta*. *The American Naturalist* **186**:716–727 DOI 10.1086/683799.

Wahli W, Dawid IB, Ryffel GU, Weber R. 1981. Vitellogenesis and the vitellogenin gene family. *Science* **212**:298–304 DOI 10.1126/science.7209528.

Wang X, Wheeler D, Avery A, Rago A, Choi J-H, Colbourne JK, Clark AG, Werren JH. 2013. Function and evolution of DNA methylation in *Nasonia vitripennis*. *PLOS Genetics* **9**:e1003872 DOI 10.1371/journal.pgen.1003872.

Wang H, Zhang H, Wong YH, Voolstra C, Ravasi T, Bajic VB, Qian P-Y. 2010. Rapid transcriptome and proteome profiling of a non-model marine invertebrate, *Bugula neritina*. *Proteomics* **10**:2972–2981 DOI 10.1002/pmic.201000056.

Weil T, Korb J, Rehli M. 2009. Comparison of queen-specific gene expression in related lower termite species. *Molecular Biology and Evolution* **26**:1841–1850 DOI 10.1093/molbev/msp095.

Weinstock GM, Robinson GE, Gibbs RA, Weinstock GM, Weinstock GM, Robinson GE, Worley KC, Evans JD, Maleszka R, Robertson HM, Weaver DB, Beye M, Bork P, Elsik CG, Evans JD, Hartfelder K, Hunt GJ, Robertson HM, Robinson GE, Maleszka R, Weinstock GM, Worley KC, Zdobnov EM, Hartfelder K, Amdam GV, Bitondi MMG, Collins AM, Cristino AS, Evans JD, Michael H, Lattorff G, Lobo CH, Moritz RFA, Nunes FMF, Page RE, Simões ZLP, Wheeler D, Carninci P, Fukuda S, Hayashizaki Y, Kai C, Kawai J, Sakazume N, Sasaki D, Tagami M, Maleszka R, Amdam GV, Albert S, Baggerman G, Beggs KT, Bloch G, Cazzamali G, Cohen M, Drapeau MD, Eisenhardt D, Emore C, Ewing MA, Fahrbach SE, Forêt S, Grimmelikhuijzen CJP, Hauser F, Hummon AB, Hunt GJ, Huybrechts J, Jones AK, Kadowaki T, Kaplan N, Kucharski R, Leboulle G, Linial M, Littleton JT, Mercer AR, Page RE, Robertson HM, Robinson GE, Richmond TA, RodriguezZas SL, Rubin EB, Sattelle DB, Schlipalius D, Schoofs L, Shemesh Y, Sweedler JV, Velarde R, Verleyen P, Vierstraete E, Williamson MR, Beye M, Ament SA, Brown SJ, Corona M, Dearden PK, Dunn WA, Elekonich MM, Elsik CG, Forêt S, Fujiyuki T, Gattermeier I, Gempe T, Hasselmann M, Kadowaki T, Kage E, Kamikouchi A, Kubo T, Kucharski R, Kunieda T, Lorenzen M, Maleszka R, Milshina NV, Morioka M, Ohashi K, Overbeek R, Page RE, Robertson HM, Robinson GE, Ross CA, Schioett M, Shippy T, Takeuchi H, Toth AL, Willis JH, Wilson MJ, Robertson HM, Zdobnov EM, Bork P, Elsik CG, Gordon KHJ, Letunic I, Hackett K, Peterson J, Felsenfeld A, Guyer M, Solignac M, Agarwala R, Cornuet JM, Elsik CG, Emore C, Hunt GJ, Monnerot M, Mougel F, Reese JT, Schlipalius D, Vautrin D, Weaver DB, Gillespie JJ, Cannone JJ, Gutell RR, Johnston JS, Elsik CG, Cazzamali G, Eisen MB, Grimmelikhuijzen CJP, Hauser F, Hummon AB, Iyer VN, Iyer V, Kosarev P, Mackey AJ, Maleszka R, Reese JT, Richmond TA, Robertson HM, Solovyev V, Souvorov A, Sweedler JV, Weinstock GM, Williamson MR, Zdobnov EM, Evans JD, Aronstein KA, Bilikova K, Chen YP, Clark AG, Decanini LI, Gelbart WM, Hetru C, Hultmark D, Imler J-L, Jiang H, Kanost M, Kimura K, Lazzaro BP, Lopez DL, Simuth J, Thompson GJ, Zou Z, De Jong P, Sodergren E, Csűrös

M, Milosavljevic A, Johnston JS, Osoegawa K, Richards S, Shu C-L, Weinstock GM, Elsik CG, Duret L, Elhaik E, Graur D, Reese JT, Robertson HM, Robertson HM, Elsik CG, Maleszka R, Weaver DB, Amdam GV, Anzola JM, Campbell KS, Childs KL, Collinge D, Crosby MA, Dickens CM, et al. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**:931–949 DOI 10.1038/nature05260.

Welch M, Lister R. 2014. Epigenomics and the control of fate, form and function in social insects. *Current Opinion in Insect Science* **1**:31–38 DOI 10.1016/j.cois.2014.04.005.

Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Beukeboom LW, Desplan C, Elsik CG, Grimmelikhuijzen CJP, Kitts P, Lynch JA, Murphy T, Oliveira DCSG, Smith CD, Zande LVD, Worley KC, Zdobnov EM, Aerts M, Albert S, Anaya VH, Anzola JM, Barchuk AR, Behura SK, Bera AN, Berenbaum MR, Bertossa RC, Bitondi MMG, Bordenstein SR, Bork P, Bornberg-Bauer E, Brunain M, Cazzamali G, Chaboub L, Chacko J, Chavez D, Childers CP, Choi J-H, Clark ME, Claudianos C, Clinton RA, Cree AG, Cristino AS, Dang PM, Darby AC, De Graaf DC, Devreese B, Dinh HH, Edwards R, Elango N, Elhaik E, Ermolaeva O, Evans JD, Foret S, Fowler GR, Gerlach D, Gibson JD, Gilbert DG, Graur D, Grunder S, Hagen DE, Han Y, Hauser F, Hultmark D, Hunter HC, Hurst GDD, Jhangian SN, Jiang H, Johnson RM, Jones AK, Junier T, Kadowaki T, Kamping A, Kapustin Y, Kechavarzi B, Kim J, Kim J, Kiryutin B, Koevoets T, Kovar CL, Kriventseva EV, Kucharski R, Lee H, Lee SL, Lees K, Lewis LR, Loehlin DW, Logsdon JM, Lopez JA, Lozado RJ, Maglott D, Maleszka R, Mayampurath A, Mazur DJ, McClure MA, Moore AD, Morgan MB, Muller J, Munoz-Torres MC, Muzny DM, Nazareth LV, Neupert S, Nguyen NB, Nunes FMF, Oakeshott JG, Okwuonu GO, Pannebakker BA, Pejaver VR, Peng Z, Pratt SC, Predel R, Pu L-L, Ranson H, Raychoudhury R, Rechtsteiner A, Reid JG, Riddle M, Romero-Severson J, Rosenberg M, Sackton TB, Sattelle DB, Schluns H, Schmitt T, Schneider M, Schuler A, Schurko AM, Shuker DM, Simoes ZLP, Sinha S, Smith Z, Souvorov A, Springauf A, Stafflinger E, Stage DE, Stanke M, Tanaka Y, Telschow A, Trent C, Vattathil S, Viljakainen L, Wanner KW, Waterhouse RM, Whitfield JB, Wilkes TE, Williamson M, Willis JH, Wolschin F, Wyder S, Yamada T, Yi SV, Zecher CN, Zhang L, Gibbs RA, Whitfield JB, Wilkes TE, Williamson M, Willis JH, Wolschin F, Wyder S, Yamada T, Yi SV, Zecher CN, Zhang L, Gibbs RA. 2010. Functional and evolutionary insights from the genomes of three parasitoid nasonia species. *Science* **327**:343–348 DOI 10.1126/science.1178028.

Wheat C. 2017. SelfBlastFilter.py. DOI 10.6084/m9.figshare.5053291.v1.

Wilson EO, Hölldobler B. 2005. The rise of the ants: a phylogenetic and ecological explanation. *Proceedings of the National Academy of Sciences of the United States of America* **102**:7411–7414 DOI 10.1073/pnas.0502264102.

Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. 2013. Mechanisms and dynamics of orphan gene emergence in insect genomes. *Genome Biology and Evolution* **5**:439–455 DOI 10.1093/gbe/evt009.

Wurm Y, Uva P, Ricci F, Wang J, Jemielity S, Iseli C, Falquet L, Keller L. 2009. Fourmidable: a database for ant genomics. *BMC Genomics* **10**:5 DOI 10.1186/1471-2164-10-5.

Wurm Y, Wang J, Keller L. 2010. Changes in reproductive roles are associated with changes in gene expression in fire ant queens. *Molecular Ecology* **19**:1200–1211 DOI 10.1111/j.1365-294X.2010.04561.x.

Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, Dijkstra MB, Oettler J, Comtesse F, Shih C-J, Wu W-J, Yang C-C, Thomas J, Beaudoing E, Praderv S, Flegel V, Cook ED, Fabbretti R, Stockinger H, Long L, Farmerie WG, Oakey J, Boomsma JJ, Pamilo P, Yi SV, Heinze J, Goodisman MAD, Farinelli L, Harshman K, Hulo N, Cerutti L, Xenarios I, Shoemaker D, Keller L. 2011. The genome of the fire ant *Solenopsis invicta*. *Proceedings of the National Academy of Sciences of the United States of America* **108**:5679–5684 DOI 10.1073/pnas.1009690108.

Wyatt TD. 2013. *Pheromones and animal behavior: chemical signals and signatures*. Second edition. Cambridge: Cambridge University Press.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam T-W, Li Y, Xu X, Wong GK-S, Wang J. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**:1660–1666 DOI 10.1093/bioinformatics/btu077.

Yan H, Simola DF, Bonasio R, Liebig J, Berger SL, Reinberg D. 2014. Eusocial insects as emerging models for behavioural epigenetics. *Nature Reviews Genetics* **15**:677–688 DOI 10.1038/nrg3787.

Zareie R, Eubel H, Millar AH, Baer B. 2013. Long-term survival of high quality sperm: insights into the sperm proteome of the honeybee *Apis mellifera*. *Journal of Proteome Research* **12**:5180–5188 DOI 10.1021/pr4004773.

Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X, Hao P. 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics* **12**:S2 DOI 10.1186/1471-2105-12-S14-S2.

Zhou X-H, Ban L-P, Iovinella I, Zhao L-J, Gao Q, Felicioli A, Sagona S, Pieraccini G, Pelosi P, Zhang L, Dani FR. 2013. Diversity, abundance, and sex-specific expression of chemosensory proteins in the reproductive organs of the locust *Locusta migratoria manilensis*. *Biological Chemistry* **394**:43–54 DOI 10.1515/hsz-2012-0114.

Zwier MV, Verhulst EC, Zwahlen RD, Beukeboom LW, Van de Zande L. 2012. DNA methylation plays a crucial role during early Nasonia development. *Insect Molecular Biology* **21**:129–138 DOI 10.1111/j.1365-2583.2011.01121.x.