Biomarkers in a socially exchanged fluid reflect colony maturity, behavior and distributed metabolism

Hakala SM[1], Meurville M-P[1], Stumpe M[2], LeBoeuf AC[1]✉

[1] Department of Biology, University of Fribourg, Chemin du Musée 10, 1700, Fribourg, Switzerland, [2] Metabolomics and Proteomics Platform, Department of Biology, University of Fribourg, Chemin du Musée 10, 1700, Fribourg, Switzerland

✉ Corresponding author: Adria LeBoeuf, email: adrialeboeuf@gmail.com

Abstract

In cooperative systems exhibiting division of labor, such as microbial communities, multicellular organisms, and social insect colonies, individual units share costs and benefits through both task specialization and exchanged materials. Socially exchanged fluids, like seminal fluid and milk, allow individuals to molecularly influence conspecifics. Many social insects have a social circulatory system, where food and endogenously produced molecules are transferred mouth-to-mouth (stomodeal trophallaxis), connecting all the individuals in the society. To understand how these endogenous molecules relate to colony life, we used quantitative proteomics to investigate the trophallactic fluid within colonies of the carpenter ant *Camponotus floridanus*. We show that different stages of the colony life cycle circulate different types of proteins: young colonies prioritize direct carbohydrate processing; mature colonies prioritize accumulation and transmission of stored resources. Further, colonies circulate proteins implicated in oxidative stress, ageing, and social insect caste determination, potentially acting as superorganismal hormones. Brood-caring individuals that are also closer to the queen in the social network (nurses) showed higher abundance of oxidative stress-related proteins. Thus, trophallaxis behavior could provide a mechanism for distributed metabolism in social insect societies. The ability to thoroughly analyze the materials exchanged between cooperative units makes social insect colonies useful models to understand the evolution and consequences of metabolic division of labor at other scales.

Key words

Ageing, Hymenoptera: Formicidae, individual variation, life history evolution, machine learning, maturation, network, population proteomics, reproduction, senescence, social behavior, social evolution, social physiology, superorganism, tradeoff

Introduction

In the course of social evolution, related organisms have formed cooperative entities such as multicellular organisms or groups of social animals (*1–3*). In social animal groups, collective decisions on movement, reproduction and even development are needed for survival (*4*, *5*). Some social groups have taken this coordination to a very high level: social insect societies develop and function as a single unit instead of as competing individuals, as 'superorganisms' paralleling the development of multicellular organisms as a single unit rather than as a set of competing cells (*6*, *7*).

In these superorganismal societies, reproductive queens and males function as the germline, and workers as the soma. Similarly to different tissues in multicellular organisms, workers can be further specialized and exhibit division of labor across different behavioral and morphological castes (*8*). While morphological castes are determined during development, the behavioral caste of an individual worker typically changes during its lifetime. At the beginning of their adult life, workers specialize inside the nest as nurses focusing on brood care, and as they age, they switch to foraging outside of the nest (*9*). Social insect colonies also go through life stages. Young colonies have an initial growth phase where they solely produce one type of worker, and only later in their life cycle they may produce more specialized worker castes and finally, males and queens (*10*). The switch to reproductive phase is a major life-history transition at the colony level, and connected to female caste determination. In social Hymenoptera, determination of whether a female larva develops into a queen or a worker, and what kind of worker exactly, is controlled by intricate differences of gene expression of the same female genome, guided primarily by environmental factors, in particular nutrition and social cues, sometimes partially influenced by genetics (*11–14*).

Coordinated function of tightly integrated groups such as social insect colonies, and subgroups such as their different castes, has been described as social physiology (*15*), consisting of various behavioral, morphological and molecular mechanisms that ensure cooperation and inclusive fitness benefits for all group members. As a part of their social physiology, some social insect societies have developed a form of social circulatory system (*16*), where nutrition and endogenously produced functional molecules, such as hormones, are transferred mouth-to-mouth from the foregut of one individual to another (*17*, *18*). This social fluid transfer is called stomodeal trophallaxis (*19*). It ensures not only that food is distributed to all adults and larvae within the colony, but also that all individuals of the colony are interconnected through shared bodily fluids. Trophallactic fluid of ants and bees typically contains endogenous proteins involved in digestion, immune defense and developmental regulation (*17*), indicating that this fluid transmits more than food.

Molecular signals are important in controlling the colony life histories and guiding caste determination both at the colony level and at the individual level. Queen pheromones are central signaling molecules acting across individuals (*20–22*). Juvenile hormone and vitellogenin are central signaling molecules in classical insect development that may also play across-individual roles in some social insects (*17*, *18*, *23*, *24*). Together with fundamental nutrient-response signaling pathways (insulin, TOR), these molecules establish the developmental trajectories of individuals (*25*, *26*). In solitary organisms, such molecules are produced and function solely within the organism's own body. In contrast, in social Hymenoptera even the molecules traditionally functioning within-individuals can be secreted to the crop and distributed among the society members through trophallaxis and the social circulatory system (*17*).

Molecular components transmitted through trophallaxis, namely juvenile hormone and juvenile hormone esterase-like proteins, have been shown to influence the development of ant larvae (*17*, *18*). Thus it is possible that molecules in trophallactic fluid may influence caste determination, similarly to honeybee workers feeding larvae with royal jelly to direct their development toward a queen fate (*17*, *27–30*). The molecular functions of trophallactic fluid are still largely unstudied, but it is known, for example, that social isolation changes its composition (*17*), with some protein components of this fluid shifting with social environment. In medicine, such correlations are typically used to define biomarkers for specific conditions and treatments, and often both accurately predict function and provide mechanistic insights (*31*). We propose that trophallactic fluid could both reflect and affect the social environments of the colony, thus providing important cues for collective decision making. However, it is not yet feasible to study the causes and consequences of the molecular composition of trophallactic fluid, as it is still largely unknown how much and what kind of qualitative and quantitative variation is present.

If indeed trophallactic fluid acts as a form of social circulatory system, managing distributed metabolic processes related to colony maturation, endogenously produced factors should correlate with colony life stages. To test this, we analyzed the trophallactic fluid proteome of the carpenter ant *Camponotus floridanus* at different scales. Our aim is to demonstrate that trophallactic fluid proteomes are filled with biomarkers reflecting biotic and abiotic conditions at both the colony and individual scale.

Results

We sought to determine whether the endogenously produced proteins present in trophallactic fluid create a robust biomarker-like signature of colony status. To assess this, we analyzed the trophallactic fluid proteomes of colonies at different stages in the colony life-cycle (*Young vs. Mature*), of colonies in natural conditions or kept in the lab (*Field vs. Lab*), and between colonies found on different nearby islands (*East vs. West*) (Figure 1, Supplemental file 1). Because trophallactic fluid proteins may be differentially expressed, transmitted, and/or sequestered across the social network of a colony, we also analyzed trophallactic fluid proteomes of single individuals in different colony 'tissues' – in-nest workers taking care of brood and out-of-nest workers (*Nurse vs. Forager*).

Overall proteome variation

Over the 73 colony and 40 single-individual trophallactic fluid samples analyzed, a total of 519 proteins were identified (Figure 2). Trophallactic fluid samples contained a set of 27 'core' trophallactic fluid proteins that were present in all samples regardless of life-cycle, life-stage or environmental conditions. Fifty-seven percent of the 519 proteins we observed were present in less than half of the samples. Even though the most common proteins displayed higher average abundance, across the entire dataset, protein abundance did not correlate with the proportion of samples containing the protein – even proteins present in only a small proportion of the samples in some cases exhibited high abundance (Figure 2 – Figure Supplement 1). The overall protein abundance was higher in colony samples relative to single individual samples, reflective of the larger trophallactic fluid volume collected. The number of proteins identified for a given sample correlated with trophallactic fluid sample volume (Pearson correlation test $p < 0.03$, $r = 0.24$ for colony samples and $p < 0.01$, $r = -0.40$ for single-individual samples).

Field-collected samples exhibited more variable proteomes than did lab-collected samples (Figure 2b, gamma GLM posthoc p-values < 0.001, Figure 2 – source data 1). Further, colonies that had been in the lab for more than one year showed less variable proteomes than did colonies that had been in the lab for only six months (Figure 2b, gamma GLM $z=-4.46$, $SE=0.04$, $p <0.001$). The trophallactic proteome variability of young and mature colonies did not differ significantly,

nor did nurses' and foragers' (Figure 2b). Foragers had fewer identified proteins in their trophallactic fluid than did nurses (Figure 2a, negative binomial z=3.72 SE=0.08 P=0.005), and there were no significant differences among the main full colony samples (Figure 2 – source data 2).

When the principal components of the trophallactic fluid proteomes were analyzed, the samples tended broadly to align with others of the same type, though clusters were not fully distinct (Figure 3A and 3B). We developed a metric, self-similarity (S), to assess the depth of difference within and across sample types (Figure 3C). Because field-collected samples had more diverse protein content even within sample types (Figure 2, 3A), the self-similarity in the *Young vs. Mature* comparison is low (Figure 3C). Single individual samples, and especially forager samples were less complex, allowing a larger proportion of their dissimilarity to be explained by sample type. Further, because our classification of nurse and forager is based on the individuals' location on brood or out-of-nest, it is possible that some nurse-classified individuals were either misclassified, transitioning from nurse to forager, or had trophallactic fluid in their crop uncharacteristic of their behavioral caste.

Comparisons of trophallactic fluid across conditions

In addition to characterizing the most abundant and core proteins of the trophallactic fluid (Figure 4, Figure 4 – Figure Supplement 1), we wanted to robustly identify proteins that differ significantly in our comparisons despite the noise inherently present in this social fluid. To accomplish this, we chose to overlay three distinct statistical approaches (Figure 1B): classical frequentist, empirical Bayes and machine-learning in the form of random forest classification. In our main comparisons, *Young vs. Mature* colonies from the field, young colonies in the *Field vs. Lab*, and individual *Nurses vs. Foragers* in the lab, we found significant differences between groups with all three analysis methods (Figure 5, Figure 5 – Figure Supplement 1, Figure 5 – Figure Supplement 2, full results for the significantly differing proteins in Supplemental File 2 and for all proteins in Supplemental Files 3-5).

For the *Young vs. Mature* comparisons, there were 10, 10 and 30 differentially abundant proteins according to frequentist t-test, empirical Bayesian LIMMA and the random forest approach, respectively. Similarly, for the *Nurse vs. Forager* comparison there were 21, 57 and 26 differentially abundant proteins, and when young colonies were brought to the laboratory and resampled after six months, 17, 31 or 29 proteins had significantly different abundance. The average accuracies of classification for comparisons with the random forest approach were: *Young vs. Mature*, 87%; *Nurse vs. Forager*, 93%; and *Field vs. Lab*, 91%. This indicates that our trained classifier can predict whether a trophallactic fluid sample originates from a nurse or a forager with 93% accuracy. We found no clear signature of spatial structure (*East vs. West*) in the trophallactic fluid proteomes. The frequentist analysis between different sampling areas found no significantly different proteins, and the random forest model did not reach high enough accuracy for this dataset to be informative (58% classification accuracy). Only the empirical Bayes approach found eight proteins that significantly differed between the sampling areas (Figure 5 – Figure Supplement 1, Supplemental Files 2 and 4).

To leverage the unique benefits of the different forms of analysis we focused our further analyses on proteins significantly different in two out of the three forms of analysis. Here, young and mature colonies differed by 12 proteins, and nurses and foragers differed by 19 proteins (Figure 5). When young colonies were brought to the laboratory and resampled six months later, the trophallactic fluid proteomes differed significantly by 20 proteins. Additionally, the single individual dataset showed that proteomes are affected both by colony of origin and by behavioral role of the individual, with 60 proteins showing significant interaction between the two factors (Supplemental File 3).

Functions of the proteins in trophallactic fluid

To investigate the functions of the proteins found in trophallactic fluid, we performed functional enrichment analysis of gene ontology terms, pathways and protein-protein interaction (PPI) networks of the trophallactic fluid proteins' *Drosophila melanogaster* orthologs. The sixty most abundant proteins in trophallactic fluid (Figure 4, Figure 4 – Figure Supplement 1) are predominantly involved in the biological processes of carbohydrate metabolism, lipid and sterol transport (Figure 6, Figure 6 – Figure Supplements 1-3, FDR < 0.00038, FDR < 0.0013 and FDR < 0.0087 respectively). The larval serum protein complex was represented by 3 out of 4 members in both the most abundant proteins and in the significantly differing proteins (hexamerins/arylphorins: Lsp1beta, Lsp1gamma and Lsp2). A strong representation of the innate immune system (Reactome pathway FDR < 6.57e-5) was evident as were lysosomal processes (KEGG pathway, FDR < 3.21e-9).

Beyond the sixty most abundant proteins in trophallactic fluid, many others are of interest as well. A critical protein in insect physiology, vitellogenin is the 93rd most abundant protein in trophallactic fluid, present in 77% and 88% of colony and single individual samples, respectively. Three of the 60 most abundant proteins had no similarity to *Drosophila* genes, and thus could not be included in the functional enrichment analysis. One of them is a putative odorant receptor, another a G-protein alpha subunit, and the third showed no orthology to characterized proteins. None of these proteins significantly differed in more than one analysis for a given comparison.

Many of the trophallactic fluid proteins, abundant or significantly differing, were represented in trophallactic fluid by multiple genes from the same protein family, in some cases part of tandem repeats in the genome, indicative of relatively recent evolution. Multiple proteins of the same family were found in the most abundant trophallactic fluid proteins (Figure 4, Figure 4 – Figure Supplement 1): a family of cathepsinD-like proteins (six in the top 60; (*17*, *32*) and a family of Maltase-B1-like proteins (five in the top 60). In the list of significantly differing proteins (Figure 5, Figure 5 – Figure Supplement 2), we observed fewer members of these families and instead saw three guanine deaminase proteins, all of which significantly differed in the *Young vs. Mature* comparison. Other families that showed duplications were glucose dehydrogenases, CREG1 and tobi-like proteins (target-of-brain-insulin).

There was an overlap of 16 proteins between the most abundant proteins and the proteins significant in two out of three of our statistical methods in any of the comparisons. The PPI network for our differentially abundant protein set (46 proteins, Figure 5) was similar to that of the most abundant proteins (Figure 4) but with increased interaction in the networks of the proteins themselves beyond what would be expected by chance (PPI enrichment p-value < 2.35e-11 in differentially abundant proteins relative to p-value < 1.59e-9 in abundant proteins), with noted enrichment in oxidation-reduction processes (FDR < 0.0026) and stronger enrichment in carbohydrate metabolic processes (FDR < 2.15e-6).

To better understand the functions of the significantly differing proteins in each comparison, we analyzed the GO terms and PPI networks of proteins significant in two out of three statistical methods separately for each of our three main comparisons (Figure 6, Figure 6 – Figure Supplements 1-3). The *Nurse vs. Forager* comparison yielded a network of proteins with more interaction than would have been predicted by chance (PPI enrichment p-value < 2.57e-4) as well as a higher degree of PPI enrichment than the other two comparisons (*Young vs. Mature* p < 0.002 and Field v Lab p < 0.02). The orthologs of differentially abundant proteins found in the behavioral caste comparison involved not only carbohydrate processing (FDR < 1.7e-4), but also oxidation-reduction and malate metabolic processes (FDR < 0.023 and FDR < 0.02, respective-

ly). These pathways have been implicated in the determination of lifespan (*33*). Indeed, two of the 46 differentially abundant proteins over all comparisons have *D. melanogaster* orthologs with the gene ontology term 'determination of adult lifespan' (Men, Sod1). The *C. floridanus* tetraspanin, significantly more abundant in nurse trophallactic fluid, is a one-to-many ortholog to the family of Tsp42E genes, one of which has also been implicated in determination of adult lifespan in *D. melanogaster*.

As trophallactic fluid samples of young and mature colonies were distinguishable by principal component analysis and our random forest classifier, we wanted to see if our trained classifier could assess a change in maturity of our young colonies after they had spent six months in the laboratory. Our random forest classifier assigned an average out-of-box maturity score to our 16 laboratory samples of 42% mature, reflecting the intermediate position of the laboratory colony samples in Figure 3.

Discussion

When an ant colony matures, the protein composition of trophallactic fluid changes in biomarker-like manner, suggesting that these proteins circulating amongst individuals play a role in age-related colony metabolism and physiology. At the individual level, certain trophallactic fluid proteins correlate with behavioral caste within the colony, a trait known to encompass both individual task requirements and age (*34*–*36*). Trophallactic fluid complexity declines over time when colonies are brought from the field to the laboratory. This may reflect dietary, microbiome or environmental complexity – typical of traits that have evolved to deal with environmental cues and stressors (e.g. immunity, (*37*)).

Overall, our data reveal a rich network of trophallactic fluid proteins connected to the principal metabolic functions of ant colonies and their life cycle. Pinpointing contexts that induce changes in trophallactic fluid, along with the exact targets and functions of the proteins, are important subjects for future work. Our establishment of biomarkers transmitted over the social circulatory system that correlate with social life will allow researchers to formulate and test hypotheses on these proteins' functional roles.

Metabolism changes with maturity

We found that trophallactic fluid includes many enzymes involved in metabolism and protein products of metabolism. Many are core trophallactic fluid proteins present in all samples, but many also differ significantly among the colony and individual life stages. Some proteins abundant in mature colonies (Lsps, apolpp (*38*–*40*)) are major insect nutrient storage proteins (*40*) that may be required to consolidate resources into large workers and sexuals, potentially acting as superorganismal hormones. Proteins abundant in foragers and young colonies (Gld, tobi, Amy, Mal, (*41*, *42*)) are well-conserved enzymes for fast sugar processing. This suggests a functional role of trophallactic fluid in the social physiology of ant colonies.

Similar shifts in protein composition or gene expression can be seen in different tissues of multicellular organisms as life-stage priorities change, for example in the midgut of drosophila females after mating, where changes in expression are observed in many genes orthologous to the proteins we found here (*42*). Additionally, *Drosophila* larval hemolymph proteome changes as development unfolds (*43*), and many of these same proteins also appear in our comparisons of worker trophallactic fluid. We suggest that regulation of larval development may at least in part occur over the social network of ants, in line with previous experimental results (*18*).

Ageing and division of metabolic labor

Viewing the colony as a superorganism, the division of reproductive labor between different types of workers (soma) and queens (germline) should result in different individuals requiring differing resources and sustaining differing metabolic costs. Our results support this hypothesis. We show that trophallactic fluid transmits numerous factors linked to ageing and coping with oxidative stress, including two of the three most well-known antioxidant enzymes: superoxide dismutase and glutathione peroxidase (*44*). These and other ageing-related proteins, such as those in redox pathways and malate metabolism (*33*, *45*), are especially elevated in nurses, the individuals that are physically the closest to the brood and queen in the trophallactic network.

These results link trophallactic fluid to one of the main topics of evolutionary ecology: the longevity-fecundity tradeoff between reproduction and coping with oxidative stress (*44*, *46*, *47*). Social insect individuals seemingly escape this tradeoff with long-lived and highly reproductive queens and short-lived, non-reproductive workers (*44*, *47*, *48*). We reveal a possible distributed metabolism which could explain why social insects seem to subvert this tradeoff. If molecules dealing with oxidative stress, or beneficial products of metabolism (nutrient storage proteins) can be spread over the circulatory system, as our results show, certain individuals may bear the costs that others in the network incur. This could account for some of the puzzling results on the plasticity of senescence in social insects (*49*–*51*), and provides a new perspective to analyze the regulatory changes of social insect reproductive castes with regard to ageing (*34*, *52*–*57*). While most previous work has focused almost exclusively on gene expression, we show that for species that engage in trophallaxis, expression studies are necessary but insufficient to understand where in the colony the relevant genes act.

Our gene-set enrichment analysis showed significant enrichment in immunity-related proteins characteristic of phagocytic hemocytes (*58*) in trophallactic fluid ('innate immune system', 'complement cascade', 'neutrophil degranulation'). These results indicate that hemocytes may themselves be transmitted mouth-to-mouth, and generally shows the involvement of the social circulatory system in colony-level immune responses with implications for social immunity. Our results do not show clear caste differentiation in the abundance of immune-related proteins, as did a study in honey bees in glands that produce trophallactic fluid proteins (*59*), though we do see similar regulation of sugar processing enzymes and glutathione-S-transferases.

Evolution of trophallactic fluid

Trophallactic fluid is one of many social fluids in biology – milk and seminal fluid are similar examples of direct transfers of biological material between individuals. Such socially exchanged materials often contain molecules that target receivers' physiology beyond the fundamental reason for the transfer (*60*, *61*), and allow social effects to directly influence the evolutionary process as indirect genetic effects (*62*–*64*). Some of the proteins we find to be significantly differing in our comparisons have previously been implicated in these other social transfers. For example, one of our protein hits is orthologous to Drosophila's CG10433, a seminal fluid protein (*65*) that impacts juvenile-hormone-associated hatch-rate post-mating (*66*). In another parallel to a phylogenetically distant social fluid, trophallactic fluid's most abundant protein is CREG1, a secreted growth-associated glycoprotein also abundant in mammalian milk (*67*). Finding molecular parallels in distinct behavioral processes hints at the fundamental role of these exchanges in the evolution of social physiology, and possibly common adaptive requirements for bioactive social fluids.

Lysosomal pathways are enriched in our most abundant trophallactic fluid proteins and in our set of significantly varying trophallactic fluid proteins between nurses and foragers, according

to the KEGG analysis. Lysosomes are acidic and can be major players in secretion, autophagic flux and exocytosis (*68–70*) – processes that may be important for nurses that feed larvae by trophallaxis. These significant lysosomal signatures we see in trophallactic fluid may indicate the mechanism of secretion (*71*), or may give us cues of how this fluid has evolved. As trophallactic fluid has become acidified in formicine ants (*72*), lysosomal genes could have been duplicated and neofunctionalized to a new role in this acidic fluid, similarly to juvenile-hormone-esterase-like proteins in trophallactic fluid (*18*). The fact that many abundant trophallactic fluid proteins represent clusters of related proteins from a few families (cathepsins, guanine deaminases, maltases) suggests there has been adaptive evolution in the proteins arriving in this fluid.

Conclusions

We show that the protein composition of ant trophallactic fluid varies across different external contexts and internal conditions both at the colony and at the individual level, suggesting that the dynamic trophallactic fluid proteome has key functions in social physiology and life cycle of colonies. By describing the natural variation of trophallactic fluid we have laid the groundwork for future studies on the possible functions of these proteins in controlling the colony life cycle, senescence and behavior.

Materials and methods

Key Resources Table				
Reagent type (species) or resource	**Designation**	**Source or reference**	**Identifiers**	**Additional information**
Other	UniProt Reference proteome (*Camponotus floridanus*); accessed February 2020	UniProt	UP000000311	
Other	NCBI RefSeq Reference proteome (*Camponotus floridanus*), v7.5	NCBI RefSeq	GCF_003227725.1	
Biological samples (*Camponotus floridanus*)	Trophallactic fluid (see details in Supplementary File 1)	This paper	Supplementary File 1	Supplementary File 1
software, algorithm	MaxQuant v1.6.2.10	MaxQuant	RRID:SCR_014485	

software, algorithm	Perseus v1.6.15.0	Perseus	RRID:SCR_015753	
software, algorithm	R 3.6.1	R	RRID:SCR_001905	
software, algorithm	Matlab 2020b	Math-works	RRID:SCR_001622	
software, algorithm	R-package MASS 7.3-53	R Project	RRID:SCR_019125	
software, algorithm	R-package LME4	R Project	RRID:SCR_015654	
software, algorithm	R-package multcomp 1.4-15	R Project	RRID:SCR_018255	
software, algorithm	LIMMA-pipeline-proteomics pipeline 3.0.0	GitHub	10.5281/zenodo.4050581	
software, algorithm	sklearn v0.22.1	Scikit-learn	RRID:SCR_019053	
software, algorithm	Python 3.7.6	Python	RRID:SCR_008394	
software, algorithm	SHapley Additive ExPlanations package v0.37.0	GitHub	RRID:SCR_021362	
software, algorithm	OMA Browser (Jan 2020 release)	OMA Browser	RRID:SCR_011978	
software, algorithm	Flybase	Flybase	RRID:SCR_006549	
software, algorithm	STRING v11	STRING	RRID:SCR_005223	

Study species

Camponotus floridanus is a common species of carpenter ant in the south-eastern USA, and has already been the focus of previous trophallactic fluid analyses (*17*, *18*). They live in dead wood or in man-made structures, often in urban habitats, and forage for honeydew, floral nectar, extra-floral nectar, and arthropod prey. Each colony has a single, singly mated queen (*73*), and polydomous nest-structures where queenless satellite nests are common. Colonies grow to tens

of thousands of workers and produce sexual brood only after multiple years of initial growth. Large established colonies have two morphologically differentiated worker castes, with variably sized small-headed minors focusing on brood care when young and foraging when old, and big-headed majors that engage in nest defense, foraging and food storage (*74*).

Colony and sample identification

The species was identified based on worker and queen morphology (*74*–*76*). In line with previous studies, we use the name *C. floridanus* with the knowledge that the taxonomy and nomenclature of the C. *atriceps* complex (to which it belongs) is not fully resolved (*74*).

We collected full young colonies (0-80 workers) and mature colony extracts (30-200 workers) on several Florida Keys islands (Figure 1 and Supplementary File 1) in winter 2019 and 2020. A colony was deemed "young" if the worker population was <100, primarily minors, and the queen was found (meaning both that the species could be clearly identified and that the nest was not a queenless satellite of an established colony), and "mature" if the colony was larger (>1000 individuals visible) and the opened nest contained many large aggressive majors. Young colonies lack majors (*77*) and individuals are generally less aggressive. We only collected mature colony samples when we also found larval brood in the opened nest. In our study area, we observed that young colonies are typically found nesting in different material than are mature colonies. Young colonies are often found under stones or in lumps of clay-like mud associated with crab burrows a short distance from the water, whereas the mature colonies were found nesting in large pieces of damp rotting wood.

Laboratory rearing

Young colonies were brought to the lab and maintained in fluon-coated plastic boxes with a mesh-ventilated lid, at 25 °C with 60 % relative humidity and a 12 h light/dark cycle. Each colony was provided with one or more glass tube for nesting, 10% sugar water, and a Bhatkar & Whitcomb diet (*78*) and some *Drosophila melanogaster*. One week prior to proteomic sampling, we substituted the honey-based food with maple syrup-based food to avoid contamination with honeybee proteins (as in LeBoeuf et al 2016).

Trophallactic fluid collection

Field samples of trophallactic fluid were collected within eight hours of ant collection. Of the 20 young colonies and 23 mature colonies, workers collected from two of the mature colonies (L and N) were subdivided into six fragments to assess variation within a single colony (two samples from major workers, two samples from brood-associated workers, and two samples from the remaining minor workers). For all other analyses only one of these for each colony (referred to as minors1) was used to avoid pseudo-replication. In the laboratory, the trophallactic fluid samples underlying the *Field vs. Lab* comparison were sampled after six months in the lab. The four colonies used for the single individual analyses had been in the lab for 18 months at the time of trophallactic fluid collection.

Trophallactic fluid was obtained from CO_2- or cold-anesthetized workers whose abdomens were gently squeezed to force them to regurgitate the contents of their crops. This method of collection was shown previously to correspond to the fluid shared during the act of adult-adult stomodeal trophallaxis (*17*). For each colony, at least 30 individuals were sampled to obtain at least 10 µl of raw trophallactic fluid. For many young colonies only smaller samples were possible, because of the low number of workers (Supplementary File 1). Young colony samples were only used for further analysis if at least 2.5 µl of trophallactic fluid were collected. For single individ-

ual samples, workers with visibly full abdomens were chosen and the obtained sample volumes ranged from 0.7 µl to 2.2 µl. An individual was classified as forager, when it was seen outside the nest tube in the feeding area of an undisturbed laboratory nest box, and a nurse, when it remained in the nest tube even after the tube was removed from the original laboratory nest and placed into a new one. For colonies from which individual samples were collected, a pooled sample was also taken from individuals that remained after individual sampling. Samples were collected with glass capillaries into 5 µl of 1 x Sigmafast Protease Inhibitor Cocktail (Sigma-Aldrich) with 50 mM Tris pH 9 in LoBind eppendorf tubes and were stored -80 C until further analysis. The total proteomics sample number is 73 colony samples of following types: 23 mature colonies with two of them sampled 6 times, 20 young colonies in the field, 16 young colonies in the laboratory, four laboratory colonies used for single individual sampling; and 40 individual samples: 20 nurses and 20 foragers.

Protein mass spectrometry sample preparation and analysis

Samples were mixed with Laemmli sample buffer and pH was adjusted with 1 M Tris-Cl, pH 7. After reduction with 1 mM DTT for 10 min at 75°C and alkylation using 5.5 mM iodoacetamide for 10 min at room temperature protein samples were separated on 4-12% gradient gels (ExpressPlus, GeneScript). Each gel lane was cut into small pieces, proteins were in-gel digested with trypsin (Promega) and the resulting peptide mixtures were processed on STAGE tips (*79*, *80*).

LC-MS/MS measurements were performed on a QExactive plus mass spectrometer (Thermo Scientific) coupled to an EasyLC 1000 nanoflow-HPLC. HPLC-column tips (fused silica) with 75 µm inner diameter were self-packed with Reprosil-Pur 120 C18-AQ, 1.9 µm (Dr. Maisch GmbH) to a length of 20 cm. A gradient of A (0.1% formic acid in water) and B (0.1% formic acid in 80% acetonitrile in water) with increasing organic proportion was used for peptide separation (loading of sample with 0% B; separation ramp: from 5-30% B within 85 min). The flow rate was 250 nl/min and for sample application 650 nl/min. The mass spectrometer was operated in the data-dependent mode and switched automatically between MS (max. of $1x10^6$ ions) and MS/MS. Each MS scan was followed by a maximum of ten MS/MS scans using normalized collision energy of 25% and a target value of 1000. Parent ions with a charge state form z = 1 and unassigned charge states were excluded from fragmentation. The mass range for MS was m/z = 370-1750. The resolution for MS was set to 70,000 and for MS/MS to 17,500. MS parameters were as follows: spray voltage 2.3 kV; no sheath and auxiliary gas flow; ion-transfer tube temperature 250°C.

The MS raw data files were uploaded into MaxQuant software (*81*), version 1.6.2.10, for peak detection, generation of peak lists of mass error corrected peptides, and for database searches. MaxQuant was set up to search both the UniProt (RRID:SCR_002380, www.uniprot.org) and NCBI (RRID:SCR_003496, www.ncbi.nlm.nih.gov) databases restricted to *C. floridanus* (UniProt, February 2020 version; NCBI RefSeq, version 7.5), along with common contaminants, such as keratins and enzymes used for digestion. Carbamidomethylcysteine was set as fixed modification and protein amino-terminal acetylation and oxidation of methionine were set as variable modifications. Three missed cleavages were allowed, enzyme specificity was trypsin/P, and the MS/MS tolerance was set to 20 ppm. The average mass precision of identified peptides was in general less than 1 ppm after recalibration. Peptide lists were further used by MaxQuant to identify and relatively quantify proteins using the following parameters: peptide and protein false discovery rates, based on a forward-reverse database, were set to 0.01, minimum peptide length was set to 7, minimum number of peptides for identification and quantitation of proteins was set to one which must be unique. The 'match-between-run' option (0.7 min) was used, which helps improve the protein identifications especially for our single-individual samples. All

proteins labelled as contaminants, reverse or only identified by site were excluded and proteins with scores less than 70 were removed. After the filtering, the dataset contained 519 proteins. Quantitative analysis was performed using iBAQ values. Intensity-based absolute quantification (iBAQ) is the quotient of sum of all identified peptides and the number of theoretically observable peptides of a protein (*82*).

Statistical analyses

Analyses of dataset characteristics were performed in Perseus v1.6.15.0 (*83*), R 3.6.1 (*84*) and Matlab R2020b (Figures 2 and 3). Differences in protein numbers among the sample types were analyzed with a negative binomial model, using the function nb.glm from the R-package MASS 7.3-53 (*85*). Proteome variability per sample type, as measured by the coefficient of variation of the iBAQ abundance of each protein when present, was analysed with a generalized linear model with gamma distribution and log-link with the R-package LME4 (1.1-26) (*86*). The package multcomp 1.4-15 was used for post-hoc testing for both models. Pearson correlation tests were used to check whether obtained protein number correlates with the sample volume. Because significant correlation was found, all further analyses were done separately for the individual samples that have small volume, and colony samples that have larger volume. Principal component analysis was run in Matlab on raw iBAQ values, for both the individual and the colony datasets.

Metric for self-similarity (S) within and across samples was calculated in Matlab2020b (https://github.com/dradri/variation2021) as follows: pairwise standardized Euclidean distances (dissimilarities, D) were calculated between each pair of samples based on square-root transformed and median subtracted protein abundances; these dissimilarities were averaged for each sample with the other samples within type$\overline{D}_{with\in}$and with the samples of the other type $\overline{D}_{across}$ and divided by the average dissimilarity to all other samples. Thus, self-similarity was calculated as:

$$S = \left| \frac{\overline{D}_{within} - \overline{D}_{across}}{\overline{D}_{all}} \right|$$

To establish the proteins whose abundance differs significantly between sample types, samples were subdivided according to three main comparisons (Figure 1): *Young vs. Mature* colonies from the field, young colonies in the *Field vs. Lab* six months later, and individual *Nurses vs. Foragers* in the lab. In addition, the extent of spatial effects was analyzed for the field-collected *Young vs. Mature* dataset by dividing the sampling locations to two areas (*East vs. West*). For the colony data, the differing sample volumes may account for a small proportion of the significant differences in the *Young vs. Mature* comparison, and to lesser extent in the *Field vs. Lab* comparison, where sample volume is collinear with the sample type. Our analyses may miss some of the proteins more abundant in the young field collected colonies which have the smallest sample volumes.

Quantitative proteomic comparisons between sample types were performed independently with three different approaches to robustly identify significantly differing proteins: 1) classical frequentist t-tests, 2) linear models with empirical Bayes variance correction, and 3) machine-learning paired with modified Shapley values. Our approach is designed to be at the same time conservative and to find most of the differing proteins among our comparisons of the trophallactic fluid. The frequentist t-tests are the most conservative, and they miss some interesting proteins due to their strict model expectations that allow only to use the most common pro-

teins. The empirical Bayes approach to cope with sample variance is a more flexible method that allows use of the entire dataset, finding important hits also among the rarer proteins, although the high amount of missing values, where iBAQ equals zero, makes the model less powerful for these proteins (*87*). The machine learning approach paired with modified Shapley values, although less well explored in the current proteomics literature, is promising for its ability to find multivariate patterns that the other methods miss, and results in interpretable classification. For each comparison, we report the full results of all three analyses in Supplemental File 2 (significantly differing proteins only) and Supplemental Files 3-5 (all results). Our results and discussion sections focus on the proteins that appear significantly different based on two out of three analysis methods (Figure 5).

Classical frequentist analysis

Within each dataset only proteins present in over 70% of the samples were analyzed. Out of an original 519 proteins, the final datasets for each comparison contained the following number of proteins: *Young vs. Mature*, 172; *Field vs. Lab*, 137; and *Nurse vs. Forager*, 136. All data were log2 transformed and median-centered, and missing data were imputed by random sampling from normal distribution with 2SD downward shift and 0.3 width for each sample. For colony datasets, we used the permutation-based FDR of 0.05, and for the single individual dataset that contained more borderline-significant proteins, we used a more stable Benjamini-Hochberg FDR with a stricter threshold of 0.01. S0 parameter (similar to fold-change) was set to 2 for all analyses. All comparisons were run as two-sample t-tests, with the *Field vs. Lab* as paired.

For the individual dataset, the combined effects of colony identity and behavioral role (*Nurse vs. Forager*) and their interaction were analyzed with two-way ANOVA, with Benjamini-Hochberg FDR corrections performed in R with the base R 3.6.1 command 'p.adjust'. Both factors were also analyzed separately with multiple- and two-sample t-tests (for colony identity and behavioral role, respectively). To allow comparison to the other statistical methods, only the simple *Nurse vs. Forager* analysis without the interaction was used for combining the lists of significantly different protein abundances. Our balanced sampling guarantees the results of this simpler model are robust enough to find the most descriptive proteins for nurse and forager trophallactic fluid, even when the more complex interactive patterns are lost.

Empirical Bayesian analysis

We implemented LIMMA (Linear Models for Microarray Data), a method for two-group comparison using empirical Bayes methods to moderate the standard errors across proteins (*87*), on our score-filtered iBAQ proteomic datasets with the LIMMA-pipeline-proteomics pipeline 3.0.0 (http://doi.org/10.5281/zenodo.4050581) developed for R 4.0.2. Data were median-normalized before comparison and all comparisons were run with a log2 fold change cutoff of 2.

Random forest and SHAP analysis

We used random forest models (sklearn.ensemble.RandomForestClassifier version 0.22.1, (*88*)) to classify samples into one of two groups for each comparison. These analyses were performed in Python 3.7.6 in a Jupyter notebook (https://github.com/dradri/variation2021). For each comparison, ten analyses were performed, each with a different seed. For each seed, the dataset was split into 80% training set and 20% test set, and a model was fit, tested and accuracy computed. If accuracy was below 85%, hyper-parameter tuning was performed with GridSearchCV (sklearn 0.22.1), and the model re-fit. A seed and its corresponding model were not retained for

further analysis if accuracy could not be improved above 75%. Accuracies for *East vs. West* ranged from 33-89% and over 20 seeds, only one could be improved above 75%. The typical parameters: max_depth, 3 or 5; max_features, 'auto'; min_samples_leaf, 3; min_samples_split, 8 or 12; n_estimators, 100 or 500. Samples were classified with out-of-box scores (Supplemental File 4). The average accuracies of classification for comparisons were: *Young vs. Mature*, 87%; *Nurse vs. Forager*, 93%; *Field vs. Lab*, 91%; *East vs. West*, 58%.

To understand which proteins contributed to the classification, we used SHAP (SHapley Additive exPlanations, shap package v0.37.0 for Python 3), a game theory tool that explains the output of machine learning models (*89*). To analyze the importance of each protein in a given comparison (feature importance), we averaged the absolute value of the Shapley values per protein across the data to derive the feature importance. Then for each protein, we averaged the feature importances over each of the ten seeded models. Proteins that have no impact on the model classification receive a feature importance value of 0. When ranked according to average feature importance, the data had an approximate Pareto distribution with an inflection point typically at feature importance of ~0.15. Thus, because there is no established cutoff for significance in this form of analysis, we chose to include as 'significant' in further analyses all proteins with a feature importance of > 0.15 (Supplemental File 5).

For random forest predictions, models trained on the classification between young and mature colonies were used to classify the same young colonies after six months in the laboratory. Out-of-box scores were averaged over five seeded models.

Orthology, gene ontology, and protein network analyses

Because little functional work has been done in ants, we analyzed gene ontology terms for the *Drosophila* orthologs to our genes of interest. Orthologs to *C. floridanus* trophallactic fluid proteins were determined with OMA ("Orthologous MAtrix" Jan 2020 release (*90*)). If no ortholog was found within OMA for a given gene, the protein sequence was protein BLASTed against *Drosophila melanogaster*. In some cases, no ortholog could be found. Annotations were compiled from NCBI RefSeq and UniProt annotations.

GO analysis was performed using both Flybase (*91*) and STRING v11 (*92*). STRING was also used for protein-protein interaction and pathway analyses, including KEGG and Reactome (SI Table 4-6). The protein-protein interaction enrichment analysis in STRING used a hypergeometric test with Benjamini-Hochberg corrected FDR. Only 43 out of the 60 most abundant proteins had sufficient annotation for use by STRING while 44 of the 46 differentially abundant proteins had sufficient annotation.

Acknowledgements and funding information

We thank Joanne Reiter for help with field work, Guillaume Kuhn for help with ant maintenance in the laboratory, and members of the Social Fluids lab and the Review Commons reviewers for their useful comments on the manuscript.

Funding: This work was supported by the Swiss National Science Foundation (PR00P3_179776 to A.C.L.), and the Swiss Government Excellence (2020.0228 to S.M.H.).

Competing interests: The authors declare no conflicts of interest.
Data and materials availability: The mass spectrometry proteomics data have been deposited

to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD028568. All other data are made available here.

Figures and legends

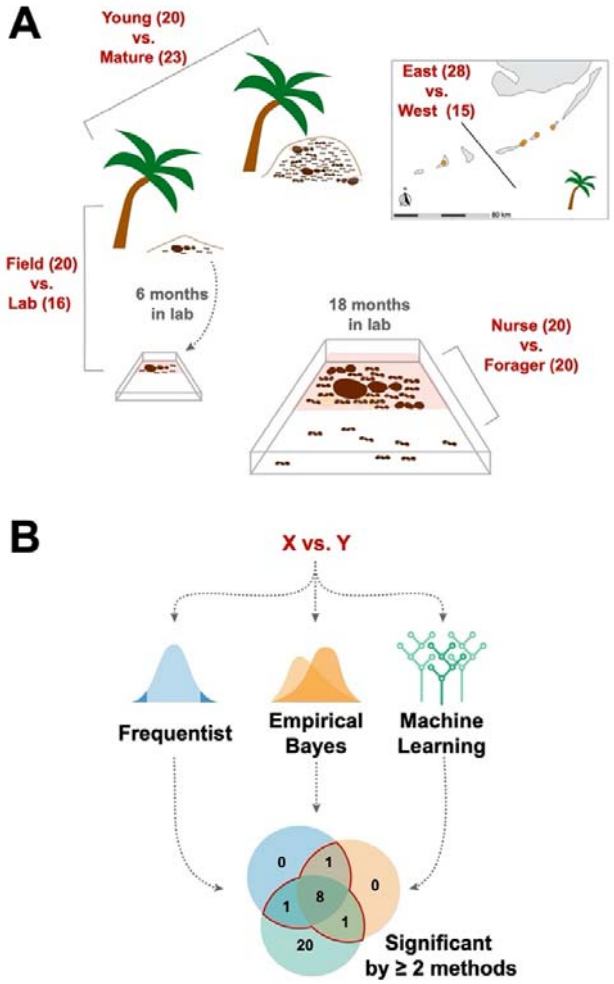

Figure 1: Schematic of study design. A. Four comparisons, *Young vs. Mature*, *Nurse vs. Forager*, *Field vs. Lab*, and *East vs. West*, analyzed in this study with sample numbers indicated in parentheses. In all comparisons sample numbers indicate colonies with the exception of *Nurse vs. Forager*, where samples are from single individuals, ten each from four colonies. Palm trees indicate field samples and boxes indicate laboratory samples. B. Schematic of analysis approach to find robustly differing proteins in each comparison. Sample information can be found in Supplementary File 1.

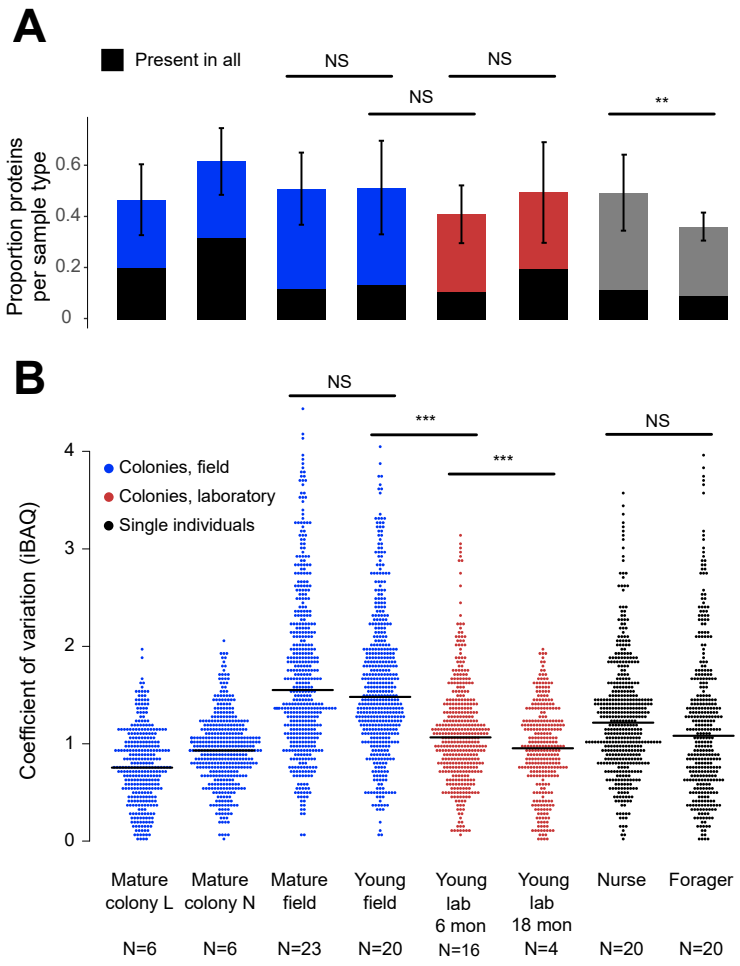

Figure 2: Protein presence in trophallactic fluid varies with biotic and abiotic factors. A. Mean ± SD of the proportion of proteins present in samples of a given type. Proportion of proteins present in all samples of a given type are highlighted in black. B. Coefficient of variation (standard deviation/mean), calculated for the iBAQ values greater than zero of all the proteins identified by sample type. Sample sizes per type are given under their names. Mature L and Mature N are mature colonies that were sampled six times to assess within-colony variation in colony samples. Significance of comparisons based on gamma GLM (A) or negative binomial GLM (B): NS indicated when $p > 0.05$ significant, ** $p < 0.01$, *** $p < 0.001$ (full results in Figure 2 – source data 1 and 2).

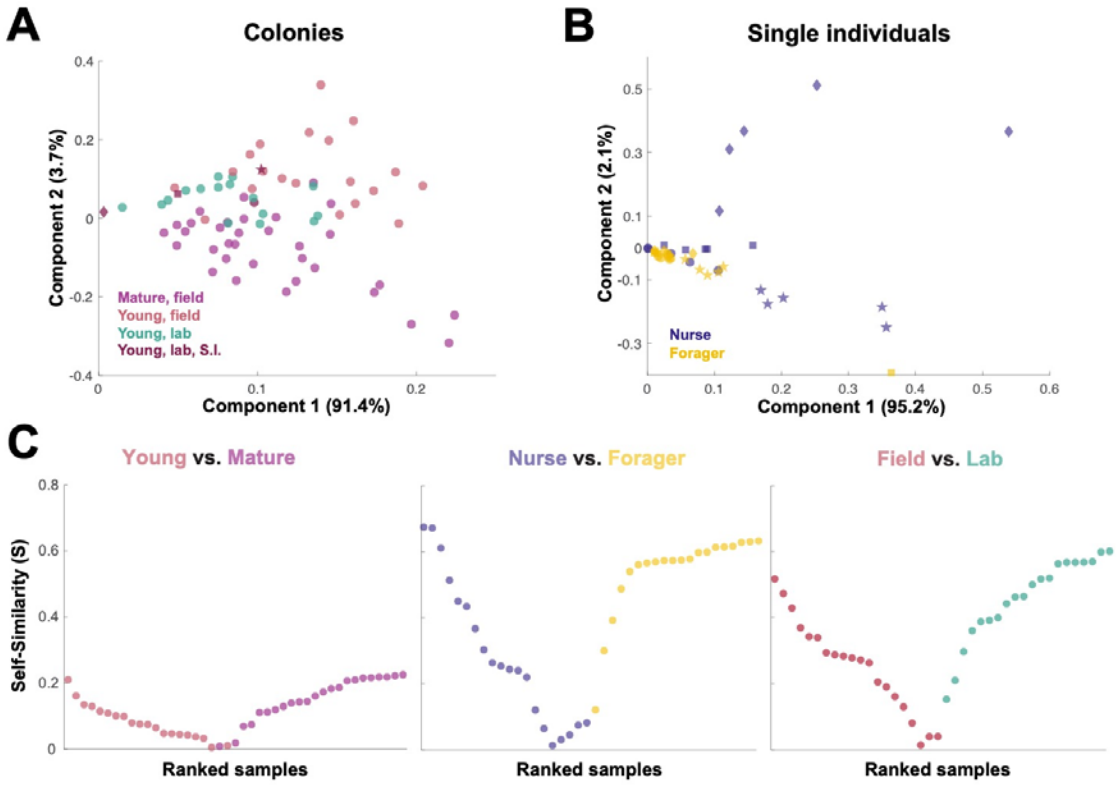

Figure 3: Similarity across trophallactic fluid proteome samples of colonies and single individuals. Principal component analysis for all proteins for (A) colony samples and (B) single individual samples from the four colonies. Symbols representing the four colonies represented in (B) can be found in maroon in (A). C. Ranked Self-similarity S for each sample type comparison. Self-similarity is the absolute value of the difference between dissimilarity within and across samples divided by the average dissimilarity of all samples (by standardized Euclidean distance of protein abundance). Samples with higher S are more similar to samples of the same type, while samples with an S of zero are equidistant to the centroids of the two sample groups.

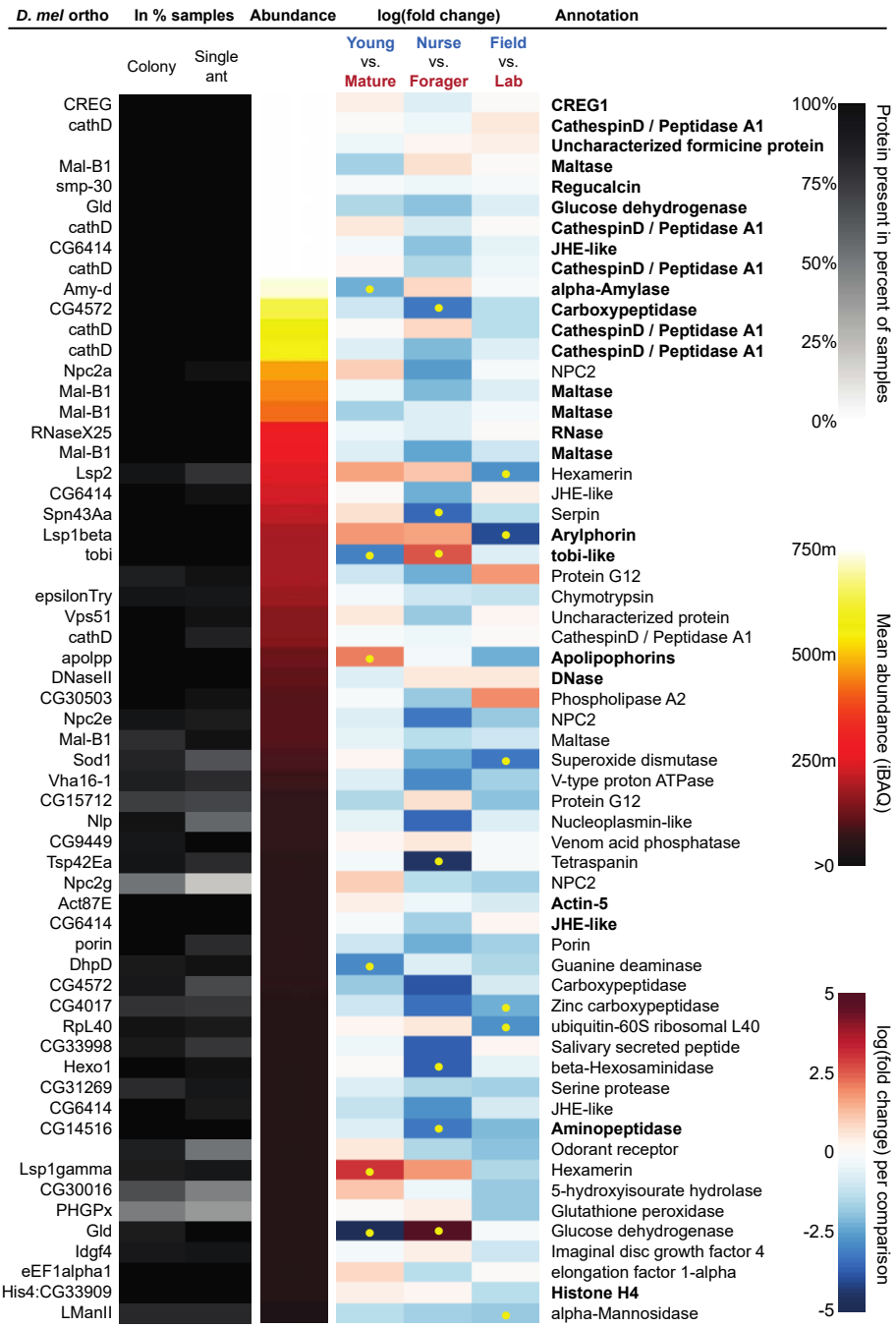

Figure 4: The sixty most abundant proteins in trophallactic fluid over 73 colony and 40 single individual samples. Ranking of abundance (including missing values). From left to right, *Drosophila melanogaster* orthologs, proportion of samples in which the protein was identified in colony samples and single individual samples, average iBAQ abundance across all samples, log2 of the fold change in abundance between types for a given comparison, the comparisons for which the protein was significant in two out of three methods are marked with yellow dots, annotation terms. Annotation terms are bolded for the 25 out of 27 core trophallactic fluid proteins that are amongst the 60 most abundant proteins. The additional but less abundant core proteins are a cathepsin (26-29-p) and a myosin heavy chain (Mhc). For protein accession numbers, see Figure 4 – Figure Supplement 1.

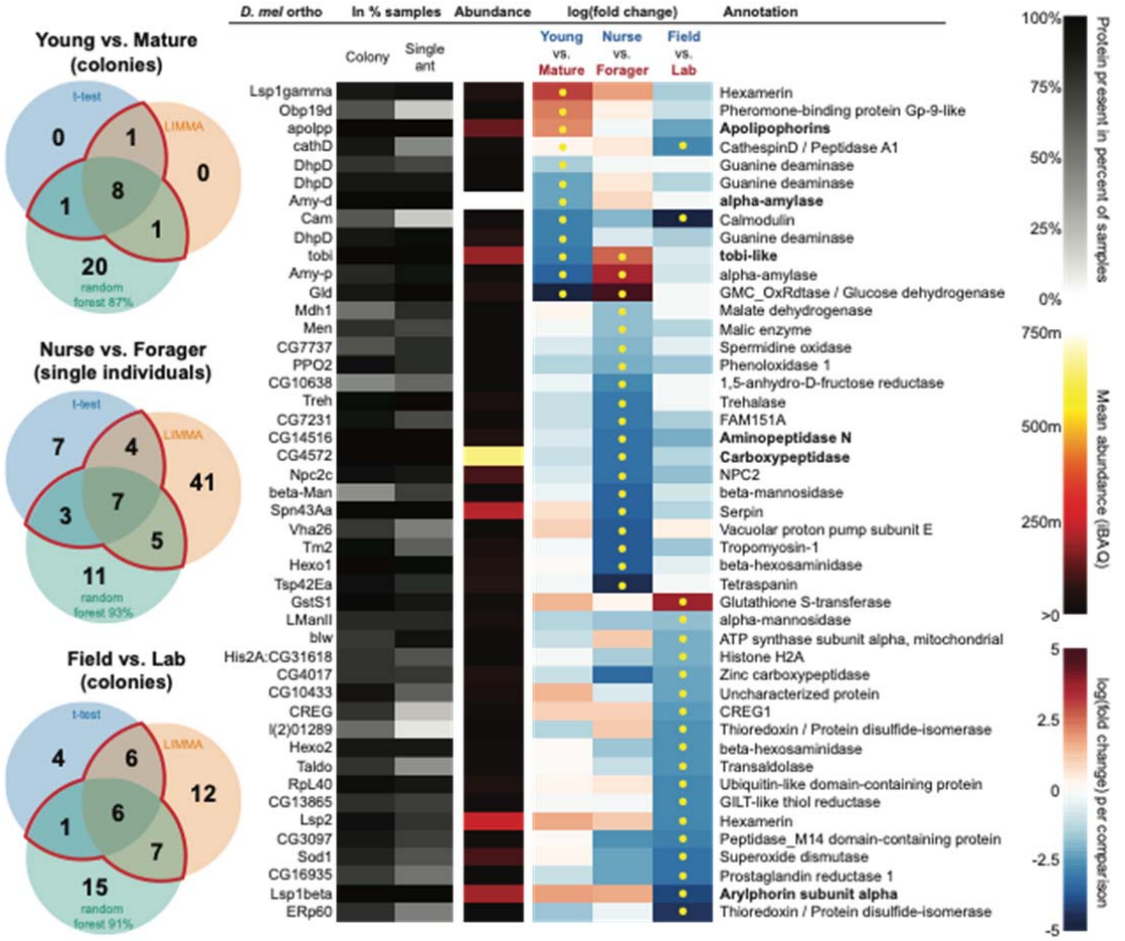

Figure 5: All proteins that significantly differ in two out of three of the analysis methods (frequentist, empirical Bayes and random forest classification with SHAP values). From left to right, Venn diagrams of significance overlap between methods, *Drosophila melanogaster* orthologs , proportion of samples in which the protein was identified in colony samples and single individual samples, average iBAQ abundance across all samples calculated without missing values, log2 of the fold change in abundance between types for a given comparison, the comparisons for which the protein was significant in two out of three methods are marked with yellow dots, annotation terms. Annotation terms are in bold for the core trophallactic fluid proteins present in all samples. For visualization of each analysis method, see Figure 5 – Figure Supplement 1. For protein accession numbers, see Figure 5 – Figure Supplement 2. For all the 135 proteins significantly differing in any analysis, see Supplemental File 2. For full model results, see Supplemental Files 3-5.

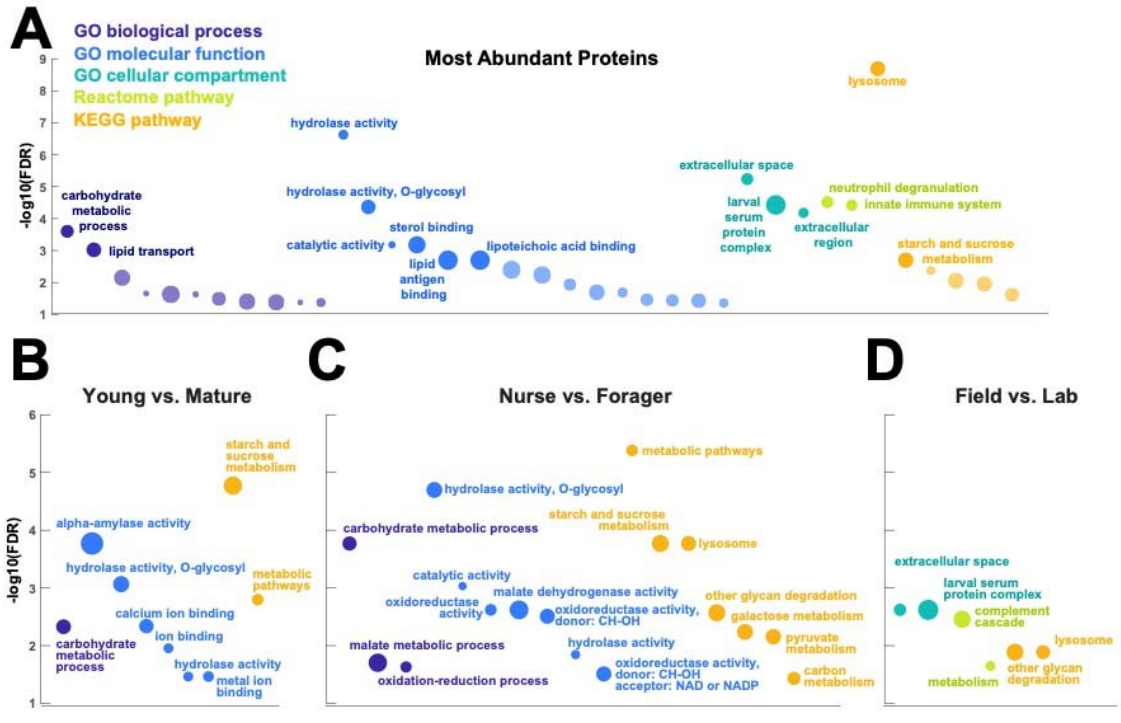

Figure 6: Gene set enrichment analysis of trophallactic fluid. Significant terms for *Drosophila melanogaster* orthologs of (A) the 60 most abundant trophallactic fluid proteins, trophallactic fluid proteins significantly differing between (B) *Young vs. Mature*, (C) *Nurse vs. Forager*, and (D) *Field vs. Lab*, with -$\log_{10}$(FDR) indicated on y-axes. Deep purple indicates GO biological process; blue, GO molecular function; turquoise, GO cellular compartment; lime green, Reactome pathway; orange, KEGG pathway. Circle size indicates strength, $\log_{10}$(observed proteins / expected proteins in a random network of this size). Full results can be found in Figure 6 – Figure Supplements 1-3.

Legends for Supplementary Figures, Files and source code

Supplemental File 1: The sampling scheme
Trophallactic fluid (TF) sampled for proteomics analysis. Date (field) indicates when the colony extract was collected from the field site, Date (TF sampling) indicates the date of the trophallactic fluid collection. Volume and ants indicate the volume collected and the number of ants collected from for each sample. The lab2019 colonies were used for single individual trophallactic fluid samples. For the two mature colonies that were sampled six times, * marks the sample that was used in the main datasets.

Figure 2 - Figure Supplement 1: Protein abundance and commonness
Protein abundances of the 519 proteins, calculated without missing values (where no matching spectra were detected), in A) the colony dataset, and B) in the single individual dataset. The proteins highlighted in red are the most abundant ones when calculated including missing values in both datasets combined, as shown in Figure 3. The red dashed line shows the cut-off used for classical frequentist statistical analyses – for the empirical Bayes and machine learning analyses all proteins were included.

Figure 2 - source data 1: Coefficient of variation by sample type
Post-hoc comparisons of gamma GLM on coefficient of variation by sample type.

Figure 2 - source data 2: Protein number by sample type
Post-hoc comparisons of negative binomial GLM on protein number explained by sample type.

Figure 3 – source code: Matlab source code to produce self-similarity scores, plots and PCA plots that make up Figure 3, https://github.com/dradri/variation2021.

Figure 3 – source data 1: Matlab MAT data file based on iBAQ values, gene names, and sample classes to produce self-similarity scores, plots and PCA plots that make up Figure 3.

Figure 4 - Figure Supplement 1: Most abundant proteins with accession numbers
The sixty most abundant proteins in trophallactic fluid over 73 colony and 40 single individual samples. Ranking of abundance included zero values. From left to right, accession numbers, proportion of samples in which the protein was identified in colony samples and single individual samples, average iBAQ abundance across all samples, $\log_2$ of the fold change in abundance between types for a given comparison, the comparisons for which the protein was significant in two out of three methods are marked with yellow dots, annotation terms. Annotation terms are bolded for the 25 out of 27 core trophallactic fluid proteins that are amongst the 60 most abundant proteins. The additional but less abundant core proteins are a cathepsin (26-29-p) and a myosin heavy chain (Mhc).

Figure 5 - Figure Supplement 1: Visualization of all results.
Venn diagrams summarizing statistical methods, frequentist volcano plots, empirical Bayes volcano plots, example SHAP value plots of feature importance the top 20 proteins. Each SHAP plot is for one of the ten models trained. For significant proteins, see Supplemental file 2, for full model results, see Supplemental files 3-5.

Figure 5 - Figure Supplement 2: Significantly differing proteins in two out of three analyses with accession numbers
All proteins that significantly differ in two out of three of the analysis methods (frequentist, empirical Bayes and random forest classification with SHAP values). From left to right, accession numbers, proportion of samples in which the protein was identified in colony samples and single individual samples, average iBAQ abundance across all samples calculated without zero values, $\log_2$ of the fold change in abundance between types for a given comparison, the

comparisons for which the protein was significant in two out of three methods are marked with yellow dots, annotation terms.

Figure 5 – source code: Jupyter notebook to run random forest analyses, https://github.com/dradri/variation2021.

Figure 6 - Figure Supplement 1: Network characteristics for all gene set enrichment analyses
The detailed results for each network are presented in Figure 6 - Figure Supplement 2 (for 60 most abundant proteins) and Figure 6 - Figure Supplement 3 (for significantly differing proteins, divided by comparison).

Figure 6 - Figure Supplement 2: Gene set enrichment analysis for the most abundant trophallactic fluid proteins
Gene set enrichment analysis results for the *D. melanogaster* orthologs of the 60 most abundant trophallactic fluid proteins. Observed gene count indicates how many proteins in the network are annotated with the term. Background gene count indicates how many proteins in total have this term, in this network and in the background. Strength describes how large the enrichment effect is: $\log_{10}$(observed proteins / expected proteins in a random network of this size). False Discovery Rate describes how significant the enrichment is. P-values are corrected for multiple testing within each category using the Benjamini–Hochberg procedure.

Figure 6 - Figure Supplement 3: Gene set enrichment analysis for the significantly differing proteins
Gene set enrichment analysis results for the *D. melanogaster* orthologs of the trophallactic fluid proteins significantly differing in two out of three of our statistical methods, first combined and then separately for the three main comparisons. Observed gene count indicates how many proteins in the network are annotated with the term. Background gene count indicates how many proteins in total have this term, in this network and in the background. Strength describes how large the enrichment effect is: $\log_{10}$(observed proteins / expected proteins in a random network of this size). False Discovery Rate describes how significant the enrichment is. P-values are corrected for multiple testing within each category using the Benjamini– Hochberg procedure. The significant annotations are indicated for GO: Biological process (GO:BP), GO: Molecular function (GO:MF), GO: Cellular component (GO:CC), Reactome pathways and KEGG pathways.

Supplemental file 2: All 135 significantly differing proteins
This supplemental file combines into a single sheet the results and additional information for all of the significantly differing proteins in our four comparisons (Young vs. Mature, Nurse vs. Forager, Field vs. Lab, East vs. West), by all of the three statistical methods (classical, empirical Bayes, machine learning). Protein accession numbers, presence in colony and individual datasets, abundance when present, fold changes by comparison and significance both by comparison and by model are shared.

Supplemental files 3-5: Full statistical results
These supplemental files share the full results of all the models run.

Supplemental file 3: Full frequentist statistical results
Statistical results for the classical frequentist models; the imputed data are also shared.

Supplemental file 4: Full empirical Bayes statistical results
For the empirical Bayes LIMMA models, results are shared as raw output tables.

Supplemental file 5: Full random forest statistical results

Accuracy, seed, and mean feature importances for each gene are reported for each model trained for the random forest analyses.

References

1. E. Szathmáry, Toward major evolutionary transitions theory 2.0. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 10104–10111 (2015).

2. D. C. Queller, A General Model for Kin Selection. *Evolution (N. Y).* **46**, 376 (1992).

3. W. D. Hamilton, The Evolution of Altruistic Behavior. *Am. Nat.* **97**, 354–356 (1963).

4. N. Miller, S. Garnier, A. T. Hartnett, I. D. Couzin, Both information and social cohesion determine collective decisions in animal groups. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5263–5268 (2013).

5. I. D. Couzin, Collective cognition in animal groups. *Trends Cogn. Sci.* **13** (2009), pp. 36–43.

6. B. R. Johnson, T. A. Linksvayer, Deconstructing the Superorganism: Social Physiology, Groundplans, and Sociogenomics. *Q. Rev. Biol.* **85**, 57–79 (2010).

7. J. J. Boomsma, R. Gawne, Superorganismality and caste differentiation as points of no return: how the major evolutionary transitions were lost in translation. *Biol. Rev.* **93**, 28–54 (2018).

8. A. F. Bourke, *Principles of social evolution* (Oxford University Press, Oxford, UK, 2011).

9. Z. Y. Huang, G. E. Robinson, Regulation of honey bee division of labor by colony age demography. *Behav. Ecol. Sociobiol.* **39**, 147–158 (1996).

10. E. O. Wilson, The insect societies (1971).

11. D. E. Wheeler, Developmental and Physiological Determinants of Caste in Social Hymenoptera : Evolutionary Implications. *Am. Nat.* **128**, 13–34 (1986).

12. K. E. Anderson, T. a Linksvayer, C. R. Smith, The causes and consequences of genetic caste determination in ants (Hymenoptera : Formicidae). *Ecology*. **11**, 119–132 (2008).

13. R. Rajakumar, S. Koch, M. Couture, M. J. Favé, A. Lillico-Ouachour, T. Chen, G. De Blasis, A. Rajakumar, D. Ouellette, E. Abouheif, Social regulation of a rudimentary organ generates complex worker-caste systems in ants. *Nature*. **562**, 574–577 (2018).

14. T. Schwander, N. Lo, M. Beekman, B. P. Oldroyd, L. Keller, Nature versus nurture in social insect caste differentiation. *Trends Ecol. Evol.* **25**, 275–82 (2010).

15. D. A. Friedman, B. R. Johnson, T. A. Linksvayer, Distributed physiology and the molecular basis of social life in eusocial insects. *Horm. Behav.* **122**, 104757 (2020).

16. W. M. Wheeler, *The Social Insects: Their Origin and Evolution* (Abingdon, Routledge, 1928).

17. A. C. LeBoeuf, P. Waridel, C. S. Brent, A. N. Gonçalves, L. Menin, D. Ortiz, O. Riba-Grognuz, A. Koto, Z. G. Soares, E. Privman, E. A. Miska, R. Benton, L. Keller, Oral transfer of chemical cues, growth proteins and hormones in social insects. *Elife*. **5**, 1–27 (2016).

18. A. C. LeBoeuf, A. B. Cohanim, C. Stoffel, C. S. Brent, P. Waridel, E. Privman, L. Keller, R. Benton, Molecular evolution of juvenile hormone esterase-like proteins in a socially exchanged fluid. *Sci. Rep.* (2018), doi:10.1038/s41598-018-36048-1.

19. M.-P. Meurville, A. LeBoeuf, Trophallaxis: the functions and evolution of social fluid exchange in ant colonies (Hymenoptera: Formicidae). *Myrmecological News*. **10**, 4136 (2021).

20. S. D. Kocher, C. M. Grozinger, Cooperation, Conflict, and the Evolution of Queen Pheromones. *J. Chem. Ecol.* **37**, 1263–1275 (2011).

21. H. F. Nijhout, D. E. Wheeler, Juvenile Hormone and the Physiological Basis of Insect Polymorphisms. *Q. Rev. Biol.* **57**, 109–133 (1982).

22. T. Pamminger, A. Buttstedt, V. Norman, A. Schierhorn, C. Botías, J. C. Jones, K. Basley, W. O. H. Hughes, The effects of juvenile hormone on Lasius niger reproduction. *J. Insect Physiol.* **95**, 1–7 (2016).

23. M. E. Scharf, C. E. Buckspan, T. L. Grzymala, X. Zhou, Regulation of polyphenic caste differentiation in the termite Reticulitermes flavipes by interaction of intrinsic and extrinsic factors. *J. Exp. Biol.* **210**, 4390–4398 (2007).

24. G. Harwood, G. Amdam, D. Freitak, The role of Vitellogenin in the transfer of immune elicitors from gut to hypopharyngeal glands in honey bees (Apis mellifera). *J. Insect Physiol.* **112**, 90–100 (2019).

25. V. Chandra, I. Fetter-Pruneda, P. R. Oxley, A. L. Ritger, S. K. McKenzie, R. Libbrecht, D. J. C. Kronauer, Social regulation of insulin signaling and the evolution of eusociality in ants. *Science (80-.).* (2018), doi:10.1126/science.aar5723.

26. R. Libbrecht, M. Corona, F. Wende, D. O. Azevedo, J. E. Serrão, L. Keller, Interplay between insulin signaling, juvenile hormone, and vitellogenin regulates maternal effects on polyphenism in ants. *Proc. Natl. Acad. Sci. U. S. A.* (2013), doi:10.1073/pnas.1221781110.

27. A. Buttstedt, R. F. A. Moritz, S. Erler, Origin and function of the major royal jelly proteins of the honeybee (Apis mellifera) as members of the yellow gene family. *Biol. Rev.* **89**, 255–269 (2014).

28. A. Buttstedt, C. H. Ihling, M. Pietzsch, R. F. A. Moritz, Royalactin is not a royal making of a queen. *Nature*, E10–E12 (2016).

29. M. Kamakura, Royalactin induces queen differentiation in honeybees. *Nature*. **473**, 478–483 (2011).

30. R. Kucharski, S. Foret, R. Maleszka, EGFR gene methylation is not involved in Royalactin controlled phenotypic polymorphism in honey bees. *Sci. Rep.* **5**, 14070 (2015).

31. K. Strimbu, J. A. Tavel, What are biomarkers? *Curr. Opin. HIV AIDS*. **5**, 463–466 (2010).

32. C. Hamilton, B. T. Lejeune, R. B. Rosengaus, Trophallaxis and prophylaxis: social immunity in the carpenter ant Camponotus pennsylvanicus. *Biol. Lett.* **7**, 89–92 (2011).

33. C. D. Wiley, J. Campisi, From Ancient Pathways to Aging Cells - Connecting Metabolism and Cellular Senescence. *Cell Metab.* **23**, 1013–1021 (2016).

34. J. Korb, K. Meusemann, D. Aumer, A. Bernadou, D. Elsner, B. Feldmeyer, S. Foitzik, J. Heinze, R. Libbrecht, M. Majoe, J. M. M. Kuhn, V. Nehring, M. A. Negroni, R. J. Paxton, A. Séguret, M. Stoldt, T. Flatt, Comparative transcriptomic analysis of the mechanisms underpinning ageing and longevity in social insects. *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* **376**, 20190728 (2021).

35. D. P. Mersch, A. Crespi, L. Keller, Tracking individuals shows spatial fidelity is a key regulator of ant social organization. *Science*. **340**, 1090–3 (2013).

36. B. Wild, D. M. Dormagen, A. Zachariae, M. L. Smith, K. S. Traynor, D. Brockmann, I. D. Couzin, T. Landgraf, Social networks predict the life and death of honey bees. *Nat. Commun.* **12**, 1–12 (2021).

37. B. P. Lazzaro, T. J. Little, Immunity in a variable world. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 15–26 (2009).

38. T. Burmester, K. Scheller, Ligands and receptors: Common theme in insect storage protein transport. *Naturwissenschaften* (1999), , doi:10.1007/s001140050656.

39. T. Burmester, Origin and evolution of arthropod hemocyanins and related proteins. *J. Comp. Physiol. B Biochem. Syst. Environ. Physiol.* (2002), , doi:10.1007/s00360-001-0247-7.

40. T. Burmester, Evolution and function of the insect hexamerins. *Eur. J. Entomol.* **96** (1999), pp. 213–225.

41. S. Buch, C. Melcher, M. Bauer, J. Katzenberger, M. J. Pankratz, Opposing Effects of Dietary Protein and Sugar Regulate a Transcriptional Target of Drosophila Insulin-like Peptide Signaling. *Cell Metab.* **7**, 321–332 (2008).

42. M. A. White, A. Bonfini, M. F. Wolfner, N. Buchon, Drosophila melanogaster sex peptide regulates mated female midgut morphology and physiology. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021), doi:10.1073/pnas.2018112118.

43. B. Handke, I. Poernbacher, S. Goetze, C. H. Ahrens, U. Omasits, F. Marty, N. Simigdala, I. Meyer, B. Wollscheid, E. Brunner, E. Hafen, C. F. Lehner, The Hemolymph Proteome of Fed and Starved Drosophila Larvae. *PLoS One*. **8**, 1–10 (2013).

44. P. Monaghan, N. B. Metcalfe, R. Torres, Oxidative stress as a mediator of life history trade-offs: Mechanisms, measurements and interpretation. *Ecol. Lett.* **12**, 75–92 (2009).

45. R. E. Koch, K. L. Buchanan, S. Casagrande, O. Crino, D. K. Dowling, G. E. Hill, W. R. Hood, M. McKenzie, M. M. Mariette, D. W. A. Noble, A. Pavlova, F. Seebacher, P. Sunnucks, E. Udino, C. R. White, K. Salin, A. Stier, Integrating Mitochondrial Aerobic Metabolism into Ecology and Evolution. *Trends Ecol. Evol.* (2021), , doi:10.1016/j.tree.2020.12.006.

46. T. Flatt, Survival costs of reproduction in Drosophila. *Exp. Gerontol.* **46**, 369–375 (2011).

47. D. Edward, T. Chapman, in *Mechanisms of life history evolution: the genetics and physiology of life history traits and trade-offs*, F. T, H. A, Eds. (Oxford, UK: Oxford University Press, 2011), pp. 137–152.

48. J. Heinze, A. Schrempf, Aging and reproduction in social insects - A mini-review. *Gerontology*. **54**, 160–167 (2008).

49. B. H. Kramer, V. Nehring, A. Buttstedt, J. Heinze, J. Korb, R. Libbrecht, K. Meusemann, R. J. Paxton, A. Séguret, F. Schaub, A. Bernadou, Oxidative stress and senescence in social insects: a significant but inconsistent link? *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20190732 (2021).

50. J. Heinze, J. Giehr, The plasticity of lifespan in social insects. *Philos. Trans. R. Soc. B Biol. Sci.* **376**, 20190734 (2021).

51. E. R. Lucas, L. Keller, The co-evolution of longevity and social life. *Funct. Ecol.* **34**, 76–87 (2020).

52. M. A. Negroni, S. Foitzik, B. Feldmeyer, Long-lived Temnothorax ant queens switch from investment in immunity to antioxidant production with age. *Sci. Rep.* (2019), doi:10.1038/s41598-019-43796-1.

53. D. Elsner, K. Meusemann, J. Korb, Longevity and transposon defense, the case of termite reproductives. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 5504–5509 (2018).

865 54. M. Corona, K. A. Hughes, D. B. Weaver, G. E. Robinson, Gene expression patterns
866 associated with queen honey bee longevity. *Mech. Ageing Dev.* **126**, 1230–1238 (2005).

867 55. C. Gstöttl, M. Stoldt, E. Jongepier, E. Bornberg-Bauer, B. Feldmeyer, J. Heinze, S. Foitzik,
868 Comparative analyses of caste, sex, and developmental stage-specific transcriptomes in
869 two Temnothorax ants. *Ecol. Evol.* **10**, 4193–4203 (2020).

870 56. M. Corona, R. Libbrecht, D. E. Wheeler, Molecular mechanisms of phenotypic plasticity in
871 social insects. *Curr. Opin. Insect Sci.* **13**, 55–60 (2016).

872 57. K. Von Wyschetzki, O. Rueppell, J. Oettler, J. Heinze, Transcriptomic signatures mirror the
873 lack of the fecundity/longevity trade-off in ant queens. *Mol. Biol. Evol.* **32**, 3173–3185
874 (2015).

875 58. U. Shokal, I. Eleftherianos, Evolution and function of thioester-containing proteins and
876 the complement system in the innate immune response. *Front. Immunol.* **8**, 1–9 (2017).

877 59. R. L. Vannette, A. Mohamed, B. R. Johnson, Forager bees (Apis mellifera) highly express
878 immune and detoxification genes in tissues associated with nectar processing. *Sci. Rep.* **5**,
879 1–9 (2015).

880 60. J. J. Bromfield, J. E. Schjenken, P. Y. Chin, A. S. Care, M. J. Jasper, S. A. Robertson, Maternal
881 tract factors contribute to paternal seminal fluid impact on metabolic phenotype in
882 offspring. *Proc. Natl. Acad. Sci. U. S. A.* (2014), doi:10.1073/pnas.1305609111.

883 61. F. Savino, S. Benetti, S. A. Liguori, M. Sorrenti, L. Cordero Di Montezemolo, Advances on
884 human milk hormones and protection against obesity. *Cell. Mol. Biol.* (2013),
885 doi:10.1170/T950.

886 62. T. A. Linksvayer, The molecular and evolutionary genetic implications of being truly
887 social for the social insects. *Adv. In Insect Phys.* **48**, 271–292 (2015).

888 63. J. W. McGlothlin, A. J. Moore, J. B. Wolf, E. D. Brodie, Interacting phenotypes and the
889 evolutionary process. III. Social evolution. *Evolution (N. Y).* (2010), doi:10.1111/j.1558-
890 5646.2010.01012.x.

891 64. J. B. Wolf, E. D. Brodie, J. M. Cheverud, A. J. Moore, M. J. Wade, Evolutionary consequences
892 of indirect genetic effects. *Trends Ecol. Evol.* (1998), doi:10.1016/S0169-5347(97)01233-
893 0.

894 65. G. D. Findlay, X. Yi, M. J. MacCoss, W. J. Swanson, Proteomics reveals novel Drosophila
895 seminal fluid proteins transferred at mating. *PLoS Biol.* (2008),
896 doi:10.1371/journal.pbio.0060178.

897 66. C. Liu, J. L. Wang, Y. Zheng, E. J. Xiong, J. J. Li, L. L. Yuan, X. Q. Yu, Y. F. Wang, Wolbachia-
898 induced paternal defect in Drosophila is likely by interaction with the juvenile hormone
899 pathway. *Insect Biochem. Mol. Biol.* **49**, 49–58 (2014).

900 67. L. Zhang, S. Boeren, J. A. Hageman, T. Van Hooijdonk, J. Vervoort, K. Hettinga, Bovine milk
901 proteome in the first 9 days: Protein interactions in maturation of the immune and
902 digestive system of the newborn. *PLoS One*. **10**, 1–19 (2015).

903 68. B. Tancini, S. Buratta, F. Delo, K. Sagini, E. Chiaradia, R. M. Pellegrino, C. Emiliani, L.
904 Urbanelli, Lysosomal exocytosis: The extracellular role of an intracellular organelle.
905 *Membranes (Basel).* **10**, 1–21 (2020).

906 69. T. Csizmadia, P. Lorincz, K. Hegedus, S. Széplaki, P. Low, G. Juhász, Molecular mechanisms
907 of developmentally programmed crinophagy in Drosophila. *J. Cell Biol.* **217**, 361–374
908 (2018).

70. T. Maruzs, Z. Simon-Vecsei, V. Kiss, T. Csizmadia, G. Juhász, On the fly: Recent progress on autophagy and aging in Drosophila. *Front. Cell Dev. Biol.* **7**, 1–15 (2019).

71. J. Martínez, I. Marmisolle, D. Tarallo, C. Quijano, Mitochondrial Bioenergetics and Dynamics in Secretion Processes. *Front. Endocrinol. (Lausanne).* **11**, 1–18 (2020).

72. S. Tragust, C. Herrmann, J. Häfner, R. Braasch, C. Tilgen, M. Hoock, M. A. Milidakis, R. Gross, H. Feldhaar, Formicine ants swallow their highly acidic poison for gut microbial selection and control. *Elife*. **9**, 1–25 (2020).

73. J. Gadau, J. Heinze, B. Hölldobler, M. Schmid, Population and colony structure of the carpenter ant Camponotus floridanus. *Mol. Ecol.* **5**, 785–792 (1996).

74. M. Deyrup, *Ants of Florida Identification and natural history* (Taylor & Francis Group, Boca Raton, FL, USA, 2017).

75. M. Deyrup, An Updated List of Florida Ants (Hymenoptera : Formicidae). *Florida Entomol.* **86**, 43–48 (2003).

76. C. S. Moreau, M. A. Deyrup, L. R. Davis, P. Oliveira, Ants of the Florida Keys: Species accounts, biogeography, and conservation (Hymenoptera: Formicidae). *J. Insect Sci.* **14**, 1–8 (2014).

77. R. L. Gibson, Soldier production in Camponotus novaeboracensis during colony growth. *Insectes Soc.* **36**, 28–41 (1989).

78. A. Bhatkar, W. H. Whitcomb, Artificial Diet for Rearing Various Species of Ants. *Florida Entomol.* **53**, 229 (1970).

79. J. Rappsilber, M. Mann, Y. Ishihama, Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).

80. A. Shevchenko, H. Tomas, J. Havlis, J. V Olsen, M. Mann, In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2006).

81. S. Tyanova, T. Temu, J. Cox, The MaxQuant computational platform for mass spectrometry – based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

82. B. Schwanhäusser, D. Busse, N. Li, Global quantification of mammalian gene expression control. *Nature*. **473**, 337–342 (2011).

83. S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, J. Cox, The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods*. **13**, 731–740 (2016).

84. R Development Core Team, A Language and Environment for Statistical Computing. *Vienna Austria R Found. Stat. Comput.* **1** (2013), p. ISBN 3-900051-07-0, , doi:10.1007/s00300-013-1395-4.

85. W. N. Venables, B. D. Ripley, *Modern Applied Statistics with S. Fourth Edition* (Springer, New York, 2002).

86. D. Bates, M. Maechler, B. Bolker, S. Walker, Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).

87. K. Kammers, R. N. Cole, C. Tiengwe, I. Ruczinski, Detecting significant changes in protein abundance. *EuPA Open Proteomics*. **7**, 11–19 (2015).

88. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

89. S. M. Lundberg, S. I. Lee, in *Advances in Neural Information Processing Systems* (2017).

90. A. M. Altenhoff, C. M. Train, K. J. Gilbert, I. Mediratta, T. Mendes de Farias, D. Moi, Y. Nevers, H. S. Radoykova, V. Rossier, A. Warwick Vesztrocy, N. M. Glover, C. Dessimoz, OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Res.* **49**, D373–D379 (2021).

91. A. Larkin, S. J. Marygold, G. Antonazzo, H. Attrill, G. Dos Santos, P. V. Garapati, J. L. Goodman, L. S. Gramates, G. Millburn, V. B. Strelets, C. J. Tabone, J. Thurmond, FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res.* **49**, D899–D907 (2021).

92. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, C. Von Mering, STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

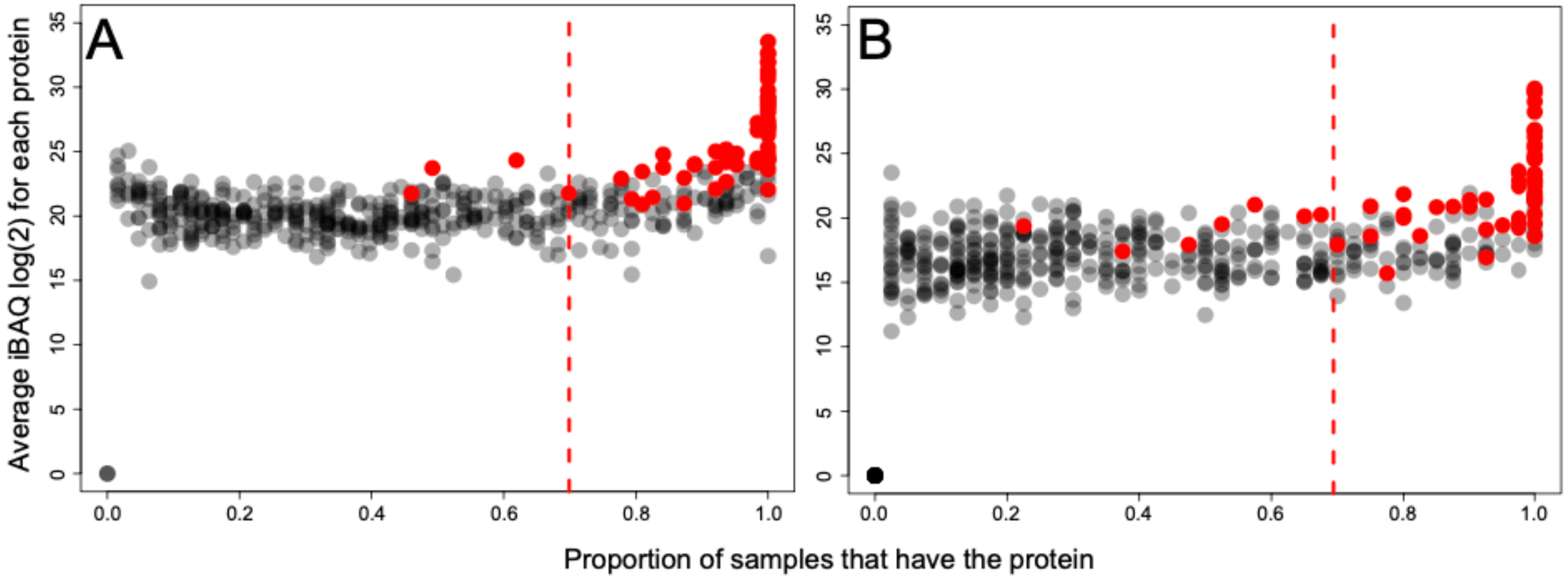

Figure 2 - Figure Supplement 1: Protein abundance and commonness

Protein abundances of the 519 proteins, calculated without missing values (where no matching spectra were detected), in A) the colony dataset, and B) in the single individual dataset. The proteins highlighted in red are the most abundant ones when calculated including missing values in both datasets combined, as shown in Figure 3. The red dashed line shows the cut-off used for classical frequentist statistical analyses – for the empirical Bayes and machine learning analyses all proteins were included.

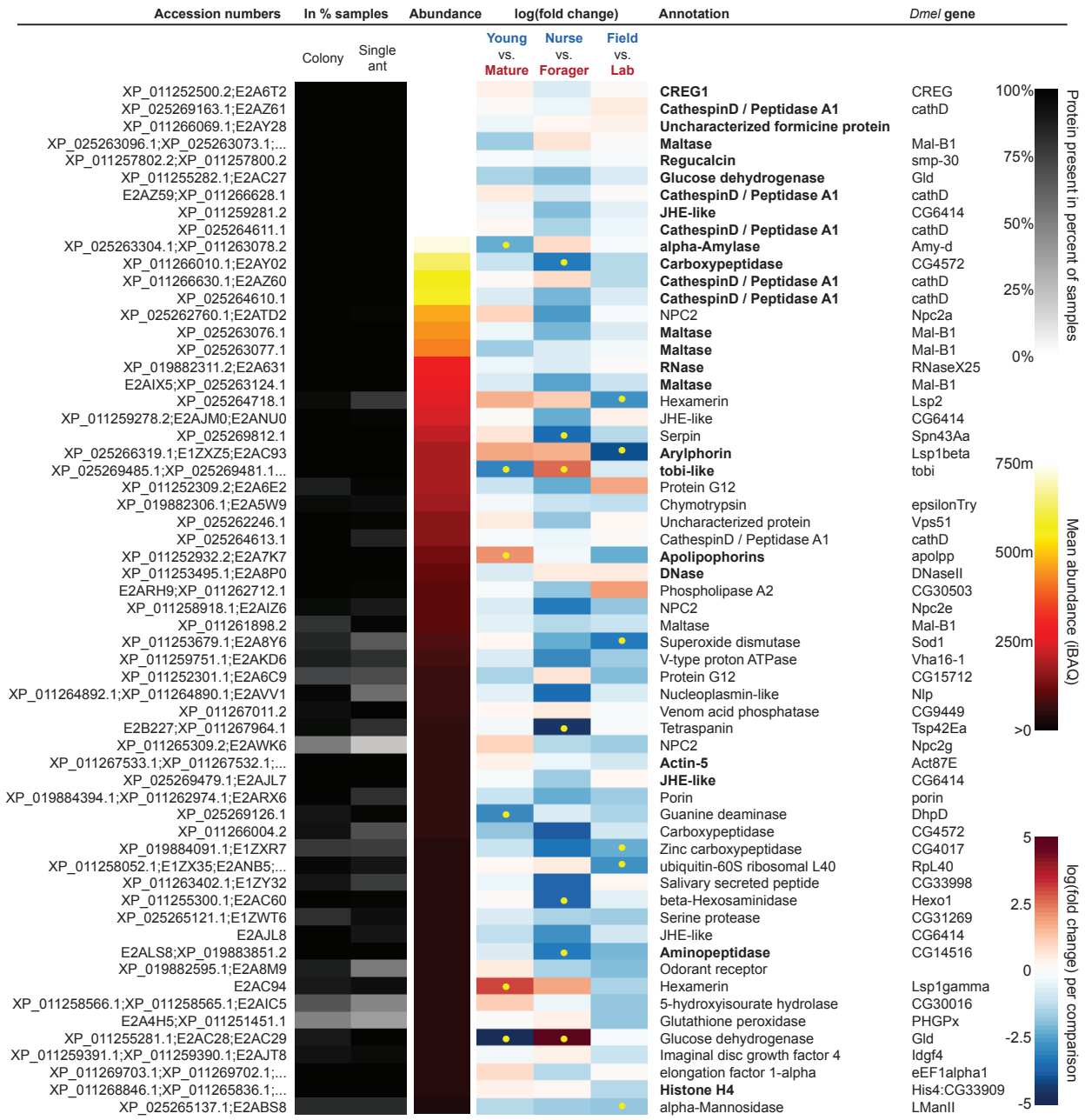

Figure 4 - Figure Supplement 1: Most abundant proteins with accession numbers

The sixty most abundant proteins in trophallactic fluid over 73 colony and 40 single individual samples. Ranking of abundance included zero values. From left to right, accession numbers, proportion of samples in which the protein was identified in colony samples and single individual samples, average iBAQ abundance across all samples, $\log_2$ of the fold change in abundance between types for a given comparison, the comparisons for which the protein was significant in two out of three methods are marked with yellow dots, annotation terms. Annotation terms are bolded for the 25 out of 27 core trophallactic fluid proteins that are amongst the 60 most abundant proteins. The additional but less abundant core proteins are a cathepsin (26-29-p) and a myosin heavy chain (Mhc).

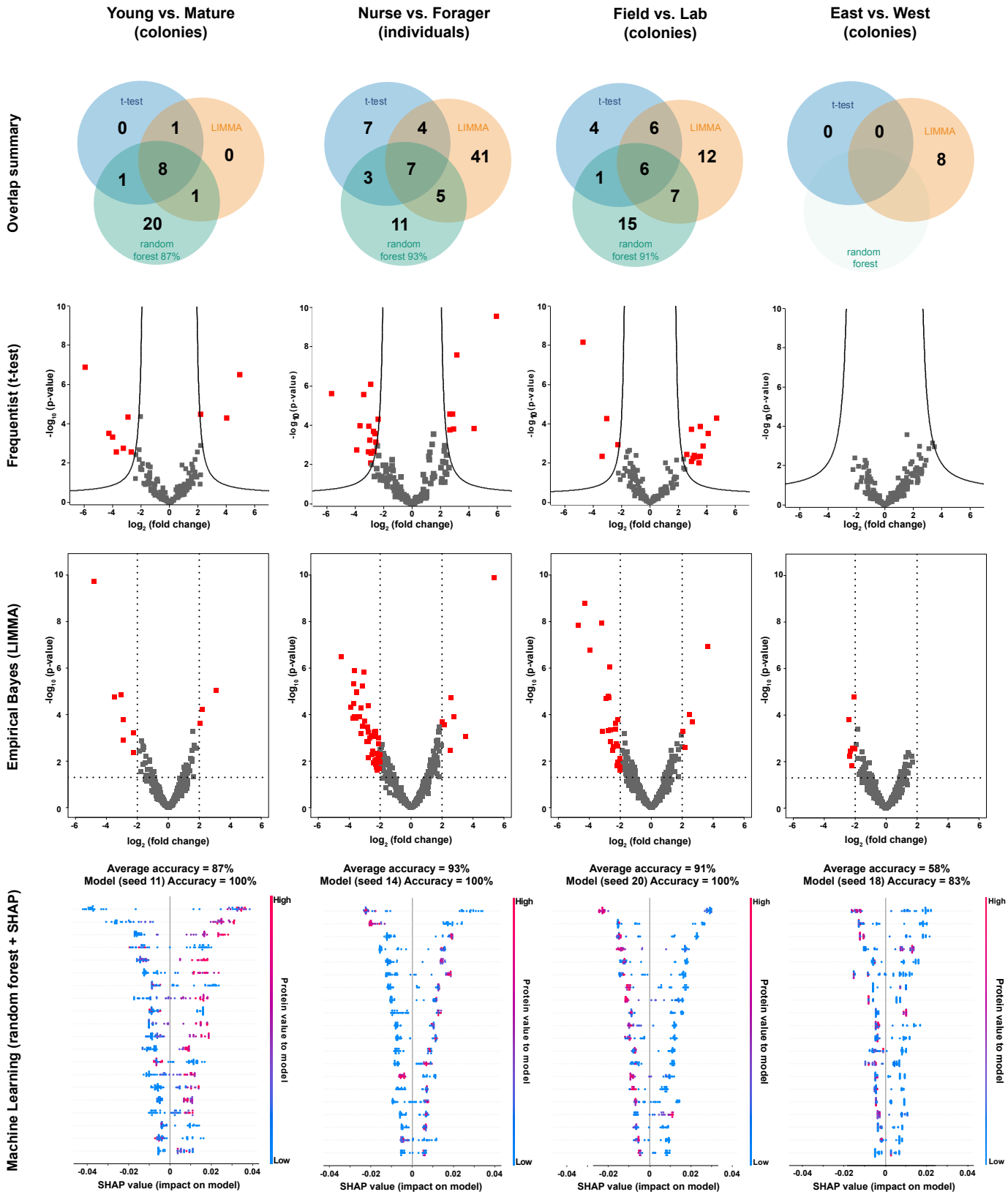

Figure 5 - Figure Supplement 1: Visualization of all results.

Venn diagrams summarizing statistical methods, frequentist volcano plots, empirical Bayes volcano plots, example SHAP value plots of feature importance the top 20 proteins. Each SHAP plot is for one of the ten models trained. For significant proteins, see Supplemental file 2, for full model results, see Supplemental files 3-5.

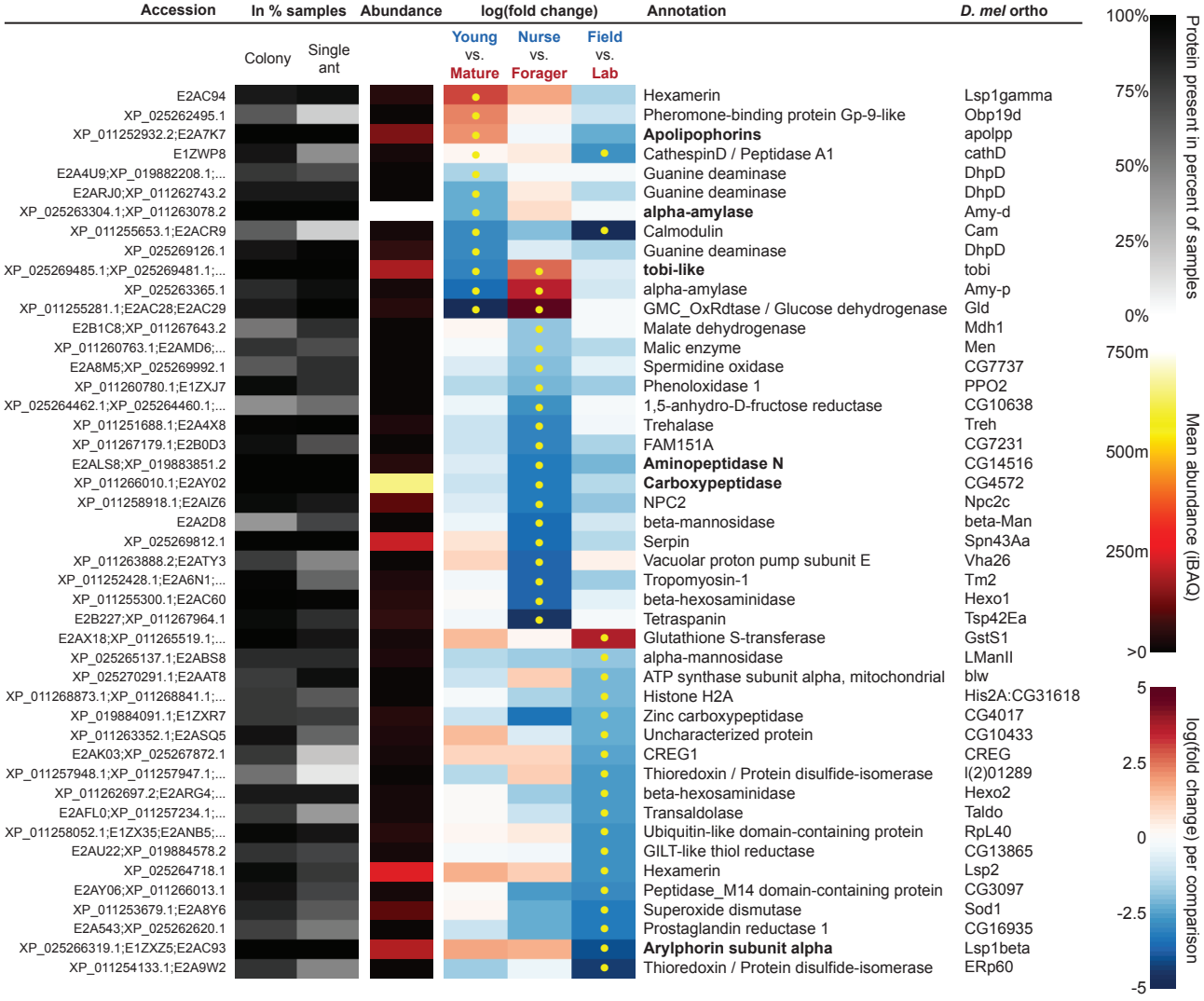

Figure 5 - Figure Supplement 2: Significantly differing proteins in two out of three analyses with accession numbers

All proteins that significantly differ in two out of three of the analysis methods (frequentist, empirical Bayes and random forest classification with SHAP values). From left to right, accession numbers, proportion of samples in which the protein was identified in colony samples and single individual samples, average iBAQ abundance across all samples calculated without zero values, $\log_2$ of the fold change in abundance between types for a given comparison, the comparisons for which the protein was significant in two out of three methods are marked with yellow dots, annotation terms.

Figure 6 - Figure Supplement 1: Network characteristics for all gene set enrichment analyses

The detailed results for each network are presented in Figure 6 - Figure Supplement 2 (for 60 most abundant proteins) and Figure 6 - Figure Supplement 3 (for significantly differing proteins, divided by comparison).

Proteins used	N nodes	N edges	Avg. local clustering coefficient	Avg. node degree:
60 most abundant	41	32	0.361	1.56
Significant in 2/3, in any comparison	44	44	0.312	2
Significant in 2/3, Nurse vs forager	19	6	0.158	0.632
Significant in 2/3, Young vs mature	10	4	0.5	0.8
Significant in 2/3, Field vs lab	20	11	0.583	1.1

Figure 6 - Figure Supplement 2: Gene set enrichment analysis for the most abundant trophallactic fluid proteins

(Next two pages) Gene set enrichment analysis results for the *D. melanogaster* orthologs of the 60 most abundant trophallactic fluid proteins. Observed gene count indicates how many proteins in the network are annotated with the term. Background gene count indicates how many proteins in total have this term, in this network and in the background. Strength describes how large the enrichment effect is: $\log_{10}$(observed proteins / expected proteins in a random network of this size). False Discovery Rate describes how significant the enrichment is. P-values are corrected for multiple testing within each category using the Benjamini–Hochberg procedure.

60 most abundant proteins							
	Description	Observed gene count	Background gene count	Strength	False discovery rate	#term ID	Matching proteins in the network
GO: Biological process	carbohydrate metabolic process	8	236	1.06	0.00025	GO:0005975	Idgf4,Hexo1,LM408,Mal-B1,Gld,smp-30,tobi,Amy-d
	lipid transport	5	75	1.36	0.00095	GO:0006869	Npc2a,Npc2g,Rfabg,Npc2e,CG30503
	sterol transport	3	22	1.67	0.0071	GO:0015918	Npc2a,Npc2g,Npc2e
	metabolic process	26	4926	0.25	0.022	GO:0008152	Idgf4,PHGPx,Hexo1,RNaseX25,RpL40,Npc2a,DhpD,CG4017,LM408,Mal-B1,Gld,Act87E,smp-30,DNaseII,CG4572,tobi,CG14516,Amy-d,epsilonTry,CG30016,cathD,His4,CG31269,CG30503,Sod,CG9449
	intracellular cholesterol transport	2	7	1.99	0.0235	GO:0032367	Npc2a,Npc2g
	primary metabolic process	23	4093	0.28	0.0235	GO:0044238	Idgf4,Hexo1,RNaseX25,RpL40,Npc2a,DhpD,CG4017,LM408,Mal-B1,Gld,Act87E,smp-30,DNaseII,CG4572,tobi,CG14516,Amy-d,epsilonTry,CG30016,cathD,His4,CG31269,CG30503
	monosaccharide metabolic process	3	55	1.27	0.0323	GO:0005996	LM408,Gld,smp-30
	nurse cell apoptotic process	2	13	1.72	0.0395	GO:0045476	DNaseII,cathD
	protein deglycosylation	2	14	1.69	0.0416	GO:0006517	Hexo1,LM408
	organic substance metabolic process	23	4432	0.25	0.0418	GO:0071704	Idgf4,Hexo1,RNaseX25,RpL40,Npc2a,DhpD,CG4017,LM408,Mal-B1,Gld,Act87E,smp-30,DNaseII,CG4572,tobi,CG14516,Amy-d,epsilonTry,CG30016,cathD,His4,CG31269,CG30503
	organic substance catabolic process	8	735	0.57	0.0423	GO:1901575	Idgf4,RNaseX25,RpL40,DhpD,DNaseII,CG4572,CG14516,cathD
	hydrolase activity	21	1685	0.63	2.37E-07	GO:0016787	CG6414,Idgf4,Hexo1,RNaseX25,DhpD,CG4017,LM408,Mal-B1,smp-30,DNaseII,CG4572,tobi,CG14516,Vha16-1,Amy-d,epsilonTry,CG30016,cathD,CG31269,CG30503,CG9449
	hydrolase activity, hydrolyzing O-glycosyl compounds	6	97	1.32	4.32E-05	GO:0004553	Idgf4,Hexo1,LM408,Mal-B1,tobi,Amy-d
	catalytic activity	24	3739	0.34	0.00066	GO:0003824	CG6414,Idgf4,PHGPx,Hexo1,RNaseX25,DhpD,CG4017,LM408,Mal-B1,Gld,smp-30,DNaseII,CG4572,tobi,CG14516,Vha16-1,Amy-d,epsilonTry,CG30016,cathD,CG31269,CG30503,Sod,CG9449
	sterol binding	3	14	1.86	0.00066	GO:0032934	Npc2a,Npc2g,Npc2e
	lipid antigen binding	2	3	2.36	0.002	GO:0030882	Npc2a,Npc2e
	lipoteichoic acid binding	2	3	2.36	0.002	GO:0070891	Npc2a,Npc2e
	lipopolysaccharide binding	2	6	2.05	0.004	GO:0001530	Npc2a,Npc2e

GO: Molecular function	endonuclease activity, active with either ribo- or deoxyribonucleic acids and producing 3'-phosphomonoesters	2	8	1.93	0.0058	GO:0016894	RNaseX25,DNaseII
	lipid binding	4	140	0.99	0.0116	GO:0008289	Npc2a,Npc2g,Rfabg,Npc2e
	peptidoglycan binding	2	18	1.58	0.0203	GO:0042834	Npc2a,Npc2e
	hydrolase activity, acting on ester bonds	6	435	0.67	0.0206	GO:0016788	CG6414,RNaseX25,smp-30,DNaseII,CG30503,CG9449
	carboxylic ester hydrolase activity	3	96	1.03	0.0345	GO:0052689	CG6414,smp-30,CG30503
	exopeptidase activity	3	100	1.01	0.0362	GO:0008238	CG4017,CG4572,CG14516
	carboxypeptidase activity	2	29	1.37	0.0372	GO:0004180	CG4017,CG4572
	peptidase activity, acting on L-amino acid peptides	6	533	0.58	0.0434	GO:0070011	CG4017,CG4572,CG14516,epsilonTry,cathD,CG31269
GO:CC	extracellular space	11	454	0.92	5.81E-06	GO:0005615	Lsp1gamma,RNaseX25,Npc2a,Lsp1beta,CG4017,Spn43Aa,Rfabg,Lsp2,Npc2e,Sod,CG9449
	larval serum protein complex	3	4	2.41	3.74E-05	GO:0005616	Lsp1gamma,Lsp1beta,Lsp2
	extracellular region	13	953	0.67	6.65E-05	GO:0005576	Idgf4,Lsp1gamma,RNaseX25,Npc2a,Lsp1beta,CG4017,smp-30,Spn43Aa,Rfabg,Lsp2,Npc2e,Sod,CG9449
Reactome	Neutrophil degranulation	10	408	0.92	3.05E-05	DME-6798695	Hexo1,RNaseX25,Npc2a,LM408,CG14516,Vha16-1,CG30016,cathD,CREG,CG9449
	Innate Immune System	11	574	0.81	3.83E-05	DME-168249	Hexo1,RNaseX25,Npc2a,CG4017,LM408,CG14516,Vha16-1,CG30016,cathD,CREG,CG9449
KEGG pathway	Lysosome	9	117	1.42	2.03E-09	dme04142	Hexo1,Npc2a,LM408,DNaseII,Npc2g,Vha16-1,Tsp42Ea,cathD,Npc2e
	Starch and sucrose metabolism	3	33	1.49	0.002	dme00500	Mal-B1,tobi,Amy-d
	Metabolic pathways	10	994	0.53	0.0042	dme01100	Hexo1,DhpD,Mal-B1,Gld,tobi,CG14516,Vha16-1,Amy-d,CG30016,CG30503
	Arachidonic acid metabolism	2	17	1.6	0.0087	dme00590	PHGPx,CG30503
	Other glycan degradation	2	22	1.49	0.0111	dme00511	Hexo1,LM408
	Galactose metabolism	2	37	1.26	0.0242	dme00052	Mal-B1,tobi

Figure 6 - Figure Supplement 3: Gene set enrichment analysis for the significantly differing proteins

(Next three pages) Gene set enrichment analysis results for the *D. melanogaster* orthologs of the trophallactic fluid proteins significantly differing in two out of three of our statistical methods, first combined and then separately for the three main comparisons. Observed gene count indicates how many proteins in the network are annotated with the term. Background gene count indicates how many proteins in total have this term, in this network and in the background. Strength describes how large the enrichment effect is: $\log_{10}$(observed proteins / expected proteins in a random network of this size). False Discovery Rate describes how significant the enrichment is. P-values are corrected for multiple testing within each category using the Benjamini– Hochberg procedure. The significant annotations are indicated for GO: Biological process (GO:BP), GO: Molecular function (GO:MF), GO: Cellular component (GO:CC), Reactome pathways and KEGG pathways.

Significant in 2 out of 3 analysis methods in any comparison							
	Description	Observed gene count	Background gene count	Strength	False discovery rate	#term ID	Matching proteins in the network
GO:BP	carbohydrate metabolic process	10	236	1.13	2.15E-06	GO:0005975	Hexo2,Treh,Tal,Hexo1,LM408,Mdh1,Gld,tobi,Amy-d,Amy-p
GO:BP	oxidation-reduction process	11	694	0.7	0.0026	GO:0055114	blw,Tal,Mdh1,Gld,Men,CG16935,CG7737,PPO2,CG10638,GstS1,Sod
GO:MF	hydrolase activity, hydrolyzing O-glycosyl compounds	8	97	1.42	1.91E-07	GO:0004553	Hexo2,Treh,Hexo1,LM408,tobi,Amy-d,Amy-p,CG12582
GO:MF	catalytic activity	27	3739	0.36	9.88E-05	GO:0003824	CG3097,Hexo2,Treh,blw,Tal,Hexo1,Vha26,DhpD,CG4017,LM408,Mdh1,Gld,Men,CG4572,tobi,CG14516,Amy-d,Amy-p,CG16935,ERp60,CG7737,PPO2,cathD,CG10638,CG12582,GstS1,Sod
GO:MF	hydrolase activity	16	1685	0.48	0.0014	GO:0016787	CG3097,Hexo2,Treh,blw,Hexo1,Vha26,DhpD,CG4017,LM408,CG4572,tobi,CG14516,Amy-d,Amy-p,cathD,CG12582
GO:MF	alpha-amylase activity	2	2	2.5	0.002	GO:0004556	Amy-d,Amy-p
GO:MF	oxidoreductase activity	9	598	0.68	0.0027	GO:0016491	Mdh1,Gld,Men,CG16935,CG7737,PPO2,CG10638,GstS1,Sod
GO:MF	carboxypeptidase activity	3	29	1.52	0.0029	GO:0004180	CG3097,CG4017,CG4572
GO:MF	beta-N-acetylglucosaminidase activity	2	4	2.2	0.0029	GO:0016231	Hexo2,Hexo1
GO:MF	exopeptidase activity	4	100	1.1	0.0045	GO:0008238	CG3097,CG4017,CG4572,CG14516
GO:MF	malate dehydrogenase activity	2	9	1.85	0.007	GO:0016615	Mdh1,Men
GO:MF	metalloexopeptidase activity	3	61	1.19	0.0132	GO:0008235	CG3097,CG4017,CG14516
GO:MF	mannosidase activity	2	14	1.66	0.0132	GO:0015923	LM408,CG12582
GO:MF	oxidoreductase activity, acting on CH-OH group of donors	3	81	1.07	0.0253	GO:0016614	Mdh1,Gld,Men
GO:MF	metallocarboxypeptidase activity	2	24	1.42	0.0307	GO:0004181	CG3097,CG4017
GO:CC	extracellular space	11	454	0.88	1.54E-05	GO:0005615	CG3097,CG10433,Lsp1gamma,Obp19d,Lsp1beta,CG4017,PPO2,Spn43Aa,Rfabg,Lsp2,Sod
GO:CC	larval serum protein complex	3	4	2.38	5.56E-05	GO:0005616	Lsp1gamma,Lsp1beta,Lsp2
Reactome pathway	Metabolism	9	833	0.53	0.0126	DME-1430728	Hexo2,Tal,Hexo1,DhpD,LM408,Mdh1,CG16935,CG12582,GstS1
Reactome pathway	Complement cascade	2	20	1.5	0.0129	DME-166658	CG3097,CG4017

KEGG pathway	Metabolic pathways	17	994	0.73	1.40E-07	dme01100	Hexo2,Treh,blw,Tal,Hexo1,Vha26,DhpD,Mdh1,Gld,Men,tobi,CG14516,Amy-d,Amy-p,CG16935,CG10638,GstS1
	Lysosome	7	117	1.28	1.97E-06	dme04142	Hexo2,Hexo1,LM408,Npc2c,Tsp42Ea,cathD,CG12582
	Other glycan degradation	4	22	1.76	1.50E-05	dme00511	Hexo2,Hexo1,LM408,CG12582
	Starch and sucrose metabolism	4	33	1.58	4.84E-05	dme00500	Treh,tobi,Amy-d,Amy-p
	Glycosphingolipid biosynthesis - ganglio series	2	5	2.1	0.0015	dme00604	Hexo2,Hexo1
	Glycosphingolipid biosynthesis - globo and isoglobo series	2	6	2.02	0.0017	dme00603	Hexo2,Hexo1
	Glycosaminoglycan degradation	2	15	1.63	0.0068	dme00531	Hexo2,Hexo1
	Pentose phosphate pathway	2	24	1.42	0.0139	dme00030	Tal,Gld
	Galactose metabolism	2	37	1.23	0.026	dme00052	tobi,CG10638
	Carbon metabolism	3	117	0.91	0.026	dme01200	Tal,Mdh1,Men
	Amino sugar and nucleotide sugar metabolism	2	47	1.13	0.0322	dme00520	Hexo2,Hexo1
	Pyruvate metabolism	2	45	1.15	0.0322	dme00620	Mdh1,Men

Significant in Nurse vs Forager

	Description	Observed gene count	Background gene count	Strength	False discovery rate	#term ID	Matching proteins in the network
GO:BP	carbohydrate metabolic process	6	236	1.27	0.00017	GO:0005975	Treh,Hexo1,Mdh1,Gld,tobi,Amy-p
	malate metabolic process	2	11	2.12	0.0198	GO:0006108	Mdh1,Men
	oxidation-reduction process	6	694	0.8	0.0234	GO:0055114	Mdh1,Gld,Men,CG7737,PPO2,CG10638
GO:MF	hydrolase activity, hydrolyzing O-glycosyl compounds	5	97	1.58	0.0000201	GO:0004553	Treh,Hexo1,tobi,Amy-p,CG12582
	catalytic activity	14	3739	0.44	0.00093	GO:0003824	Treh,Hexo1,Vha26,Mdh1,Gld,Men,CG4572,tobi,CG14516,Amy-p,CG7737,PPO2,CG10638,CG12582
	oxidoreductase activity	6	598	0.87	0.0024	GO:0016491	Mdh1,Gld,Men,CG7737,PPO2,CG10638
	malate dehydrogenase activity	2	9	2.21	0.0024	GO:0016615	Mdh1,Men
	oxidoreductase activity, acting on CH-OH group of donors	3	81	1.43	0.0031	GO:0016614	Mdh1,Gld,Men
	hydrolase activity	8	1685	0.54	0.0143	GO:0016787	Treh,Hexo1,Vha26,CG4572,tobi,CG14516,Amy-p,CG12582

	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	2	53	1.44	0.0309	GO:0016616	Mdh1,Men
KEGG pathways	Metabolic pathways	10	994	0.87	0.00000416	dme01100	Treh,Hexo1,Vha26,Mdh1,Gld,Men,tobi,CG14516,Amy-p,CG10638
	Starch and sucrose metabolism	3	33	1.82	0.00017	dme00500	Treh,tobi,Amy-p
	Lysosome	4	117	1.4	0.00017	dme04142	Hexo1,Npc2c,Tsp42Ea,CG12582
	Other glycan degradation	2	22	1.82	0.0027	dme00511	Hexo1,CG12582
	Galactose metabolism	2	37	1.6	0.0058	dme00052	tobi,CG10638
	Pyruvate metabolism	2	45	1.51	0.007	dme00620	Mdh1,Men
	Carbon metabolism	2	117	1.1	0.0369	dme01200	Mdh1,Men
Significant in Young vs Mature							
GO:BP	carbohydrate metabolic process	4	236	1.37	0.0047	GO:0005975	Gld,tobi,Amy-d,Amy-p
GO:MF	alpha-amylase activity	2	2	3.14	0.00017	GO:0004556	Amy-d,Amy-p
	hydrolase activity, hydrolyzing O-glycosyl compounds	3	97	1.63	0.00086	GO:0004553	tobi,Amy-d,Amy-p
	calcium ion binding	3	206	1.31	0.0046	GO:0005509	Amy-d,Amy-p,Cam
	ion binding	6	1993	0.62	0.0111	GO:0043167	DhpD,Gld,Amy-d,Amy-p,Rfabg,Cam
	hydrolase activity	5	1685	0.62	0.0343	GO:0016787	DhpD,tobi,Amy-d,Amy-p,cathD
	metal ion binding	4	1027	0.73	0.0343	GO:0046872	DhpD,Amy-d,Amy-p,Cam
KEGG	Starch and sucrose metabolism	3	33	2.1	0.0000169	dme00500	tobi,Amy-d,Amy-p
	Metabolic pathways	5	994	0.85	0.0016	dme01100	DhpD,Gld,tobi,Amy-d,Amy-p
Significant in Field vs Lab							
GO:CC	extracellular space	6	454	0.96	0.0024	GO:0005615	CG3097,CG10433,Lsp1beta,CG4017,Lsp2,Sod
	larval serum protein complex	2	4	2.54	0.0024	GO:0005616	Lsp1beta,Lsp2
Reactome	Complement cascade	2	20	1.84	0.0035	DME-166658	CG3097,CG4017
	Metabolism	5	833	0.62	0.0225	DME-1430728	Hexo2,Tal,LM408,CG16935,GstS1
KEGG	Other glycan degradation	2	22	1.8	0.0127	dme00511	Hexo2,LM408
	Lysosome	3	117	1.25	0.0127	dme04142	Hexo2,LM408,cathD