RESEARCH ARTICLE

Figures and figure supplements

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

Sanja M Hakala *et al*

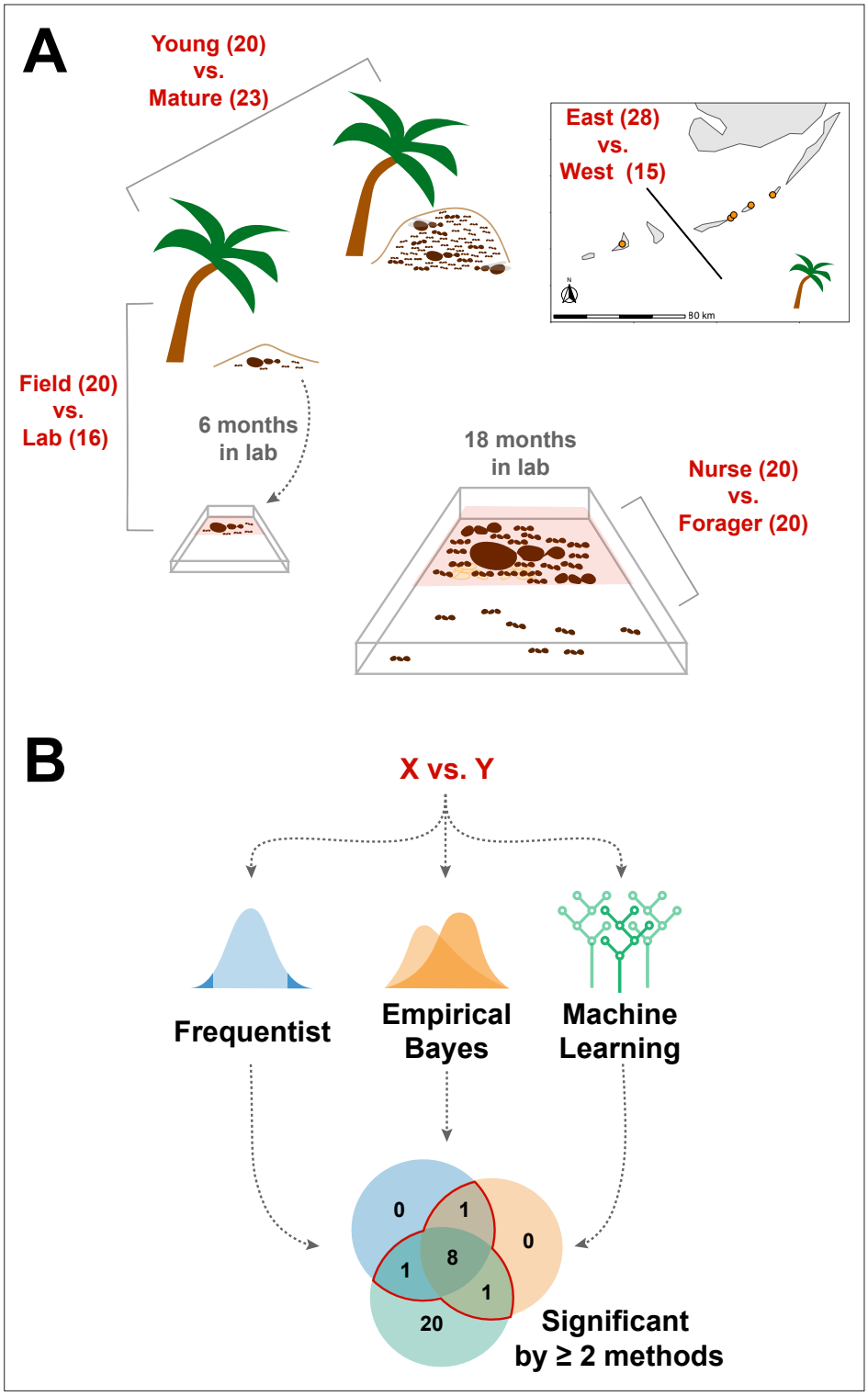

Figure 1. Schematic of study design. (**A**) Four comparisons, *Young vs. Mature*, *Nurse vs. Forager*, *Field vs. Lab*, and *East vs. West*, analyzed in this study with sample numbers indicated in parentheses. In all comparisons sample numbers indicate colonies with the exception of *Nurse vs. Forager*, 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***.

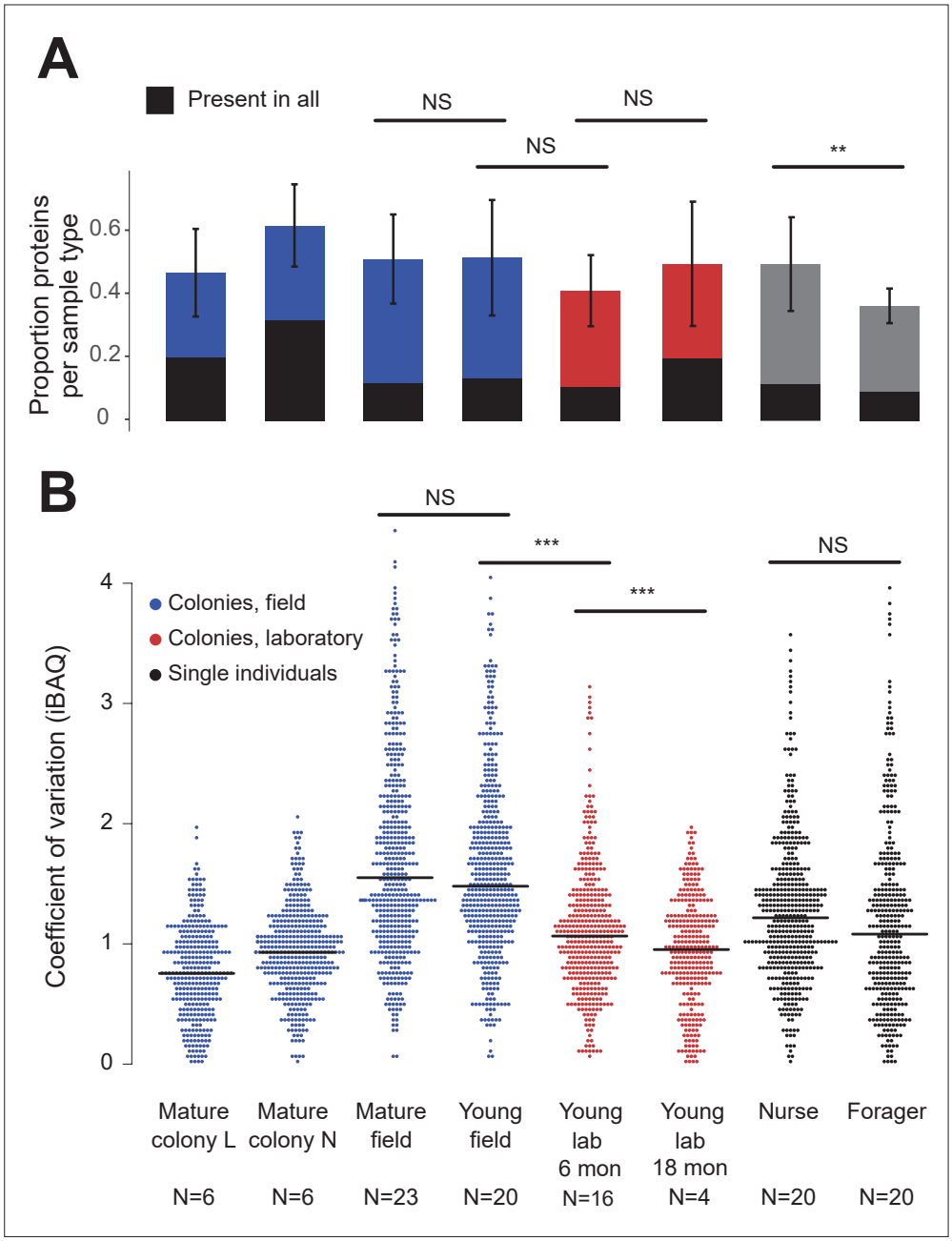

Figure 2. Protein presence in trophallactic fluid varies with biotic and abiotic factors. (**A**) Mean ± SD of the proportion of proteins present in samples of a given type. Proportion of proteins present in all samples of a given type are highlighted in black. (**B**) Coefficient of variation (standard 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***; ***Figure 2—source data 2***).

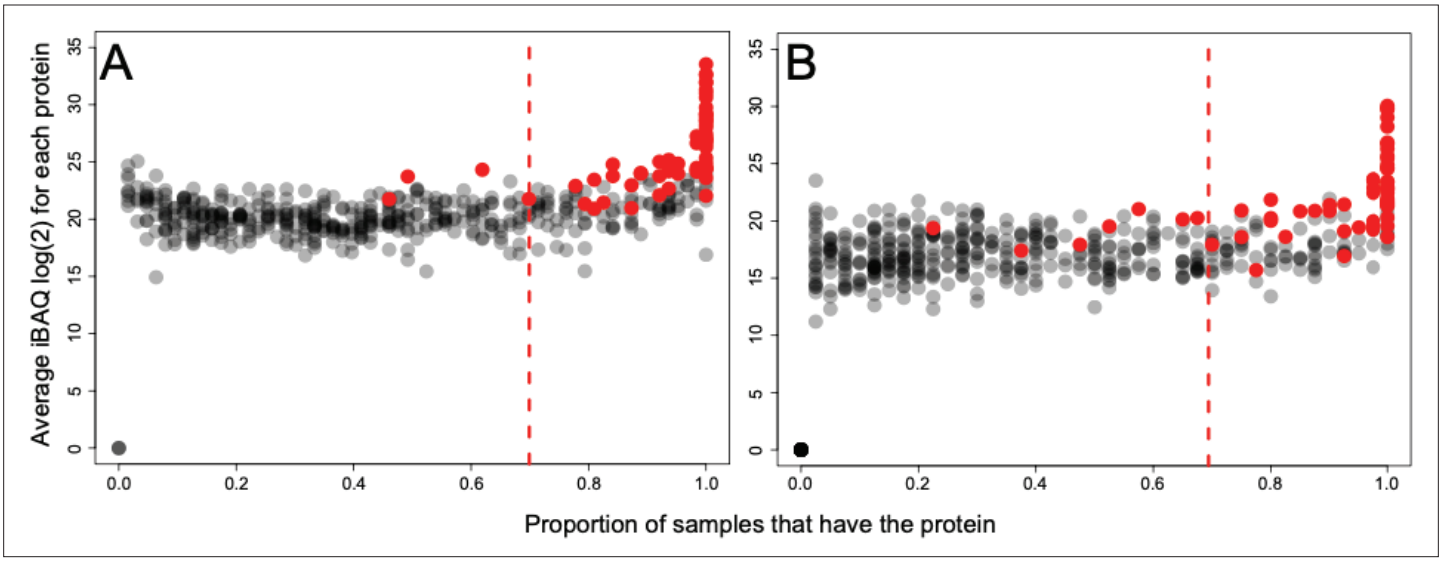

Figure 2—figure supplement 1. Protein abundance and commonness. Protein abundances of the 519 proteins, calculated without missing values (where no matching spectra were detected), in (**A**) the colony dataset, and (**B**) in the single individual dataset. The proteins highlighted in red are the most 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.

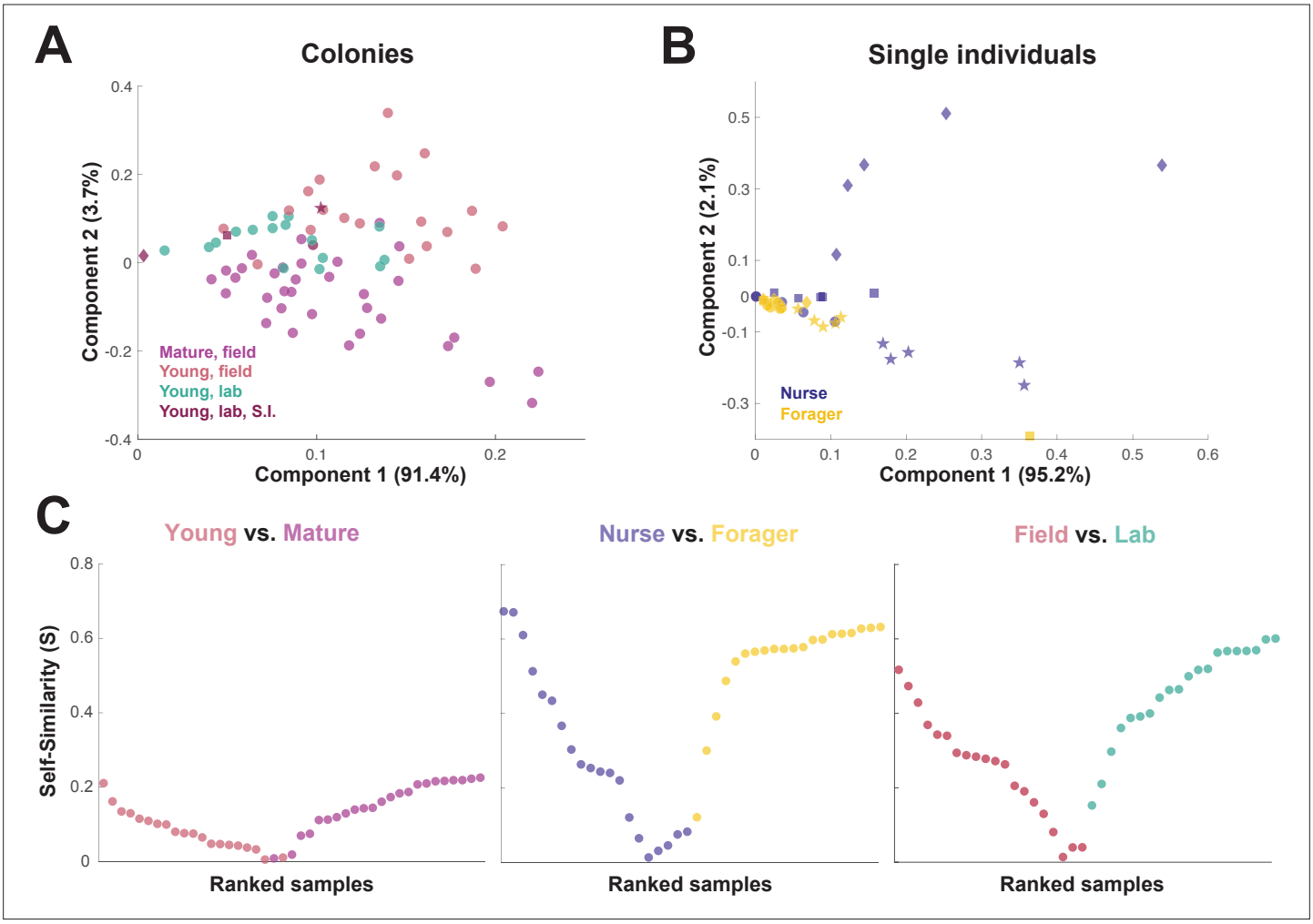

Figure 3. Similarity across trophallactic fluid proteome samples of colonies and single individuals. Principal component analysis for all proteins for (**A**) colony samples and (**B**) single individual samples from the four colonies. Symbols representing the four colonies represented in (**B**) can 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.

NO_FIGURE_FOUND**Figure 3—source code 1.** Matlab source code to produce self-similarity scores, plots and PCA plots that make up ***Figure 3***, https://github.com/dradri/variation2021.Selfsimilaritycode.mlx

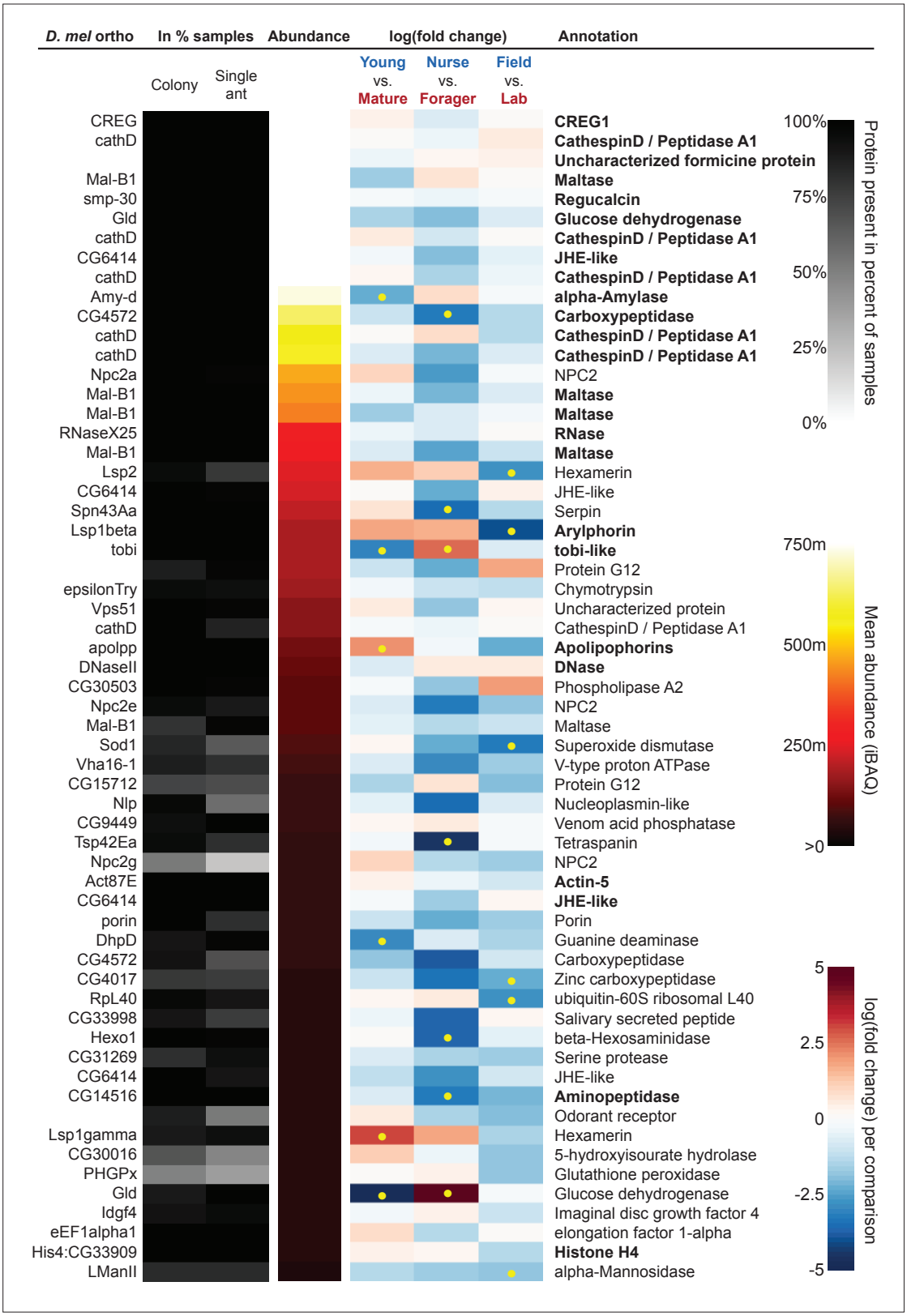

Figure 4. The sixty most abundant proteins in trophallactic fluid over 73 colony and 40 single individual samples. Ranking of abundance (including missing values). From left to right, *Drosophila melanogaster* orthologs, proportion of samples in which the protein was identified in colony samples and 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

Figure 4 continued on next page

Figure 4 continued

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***.

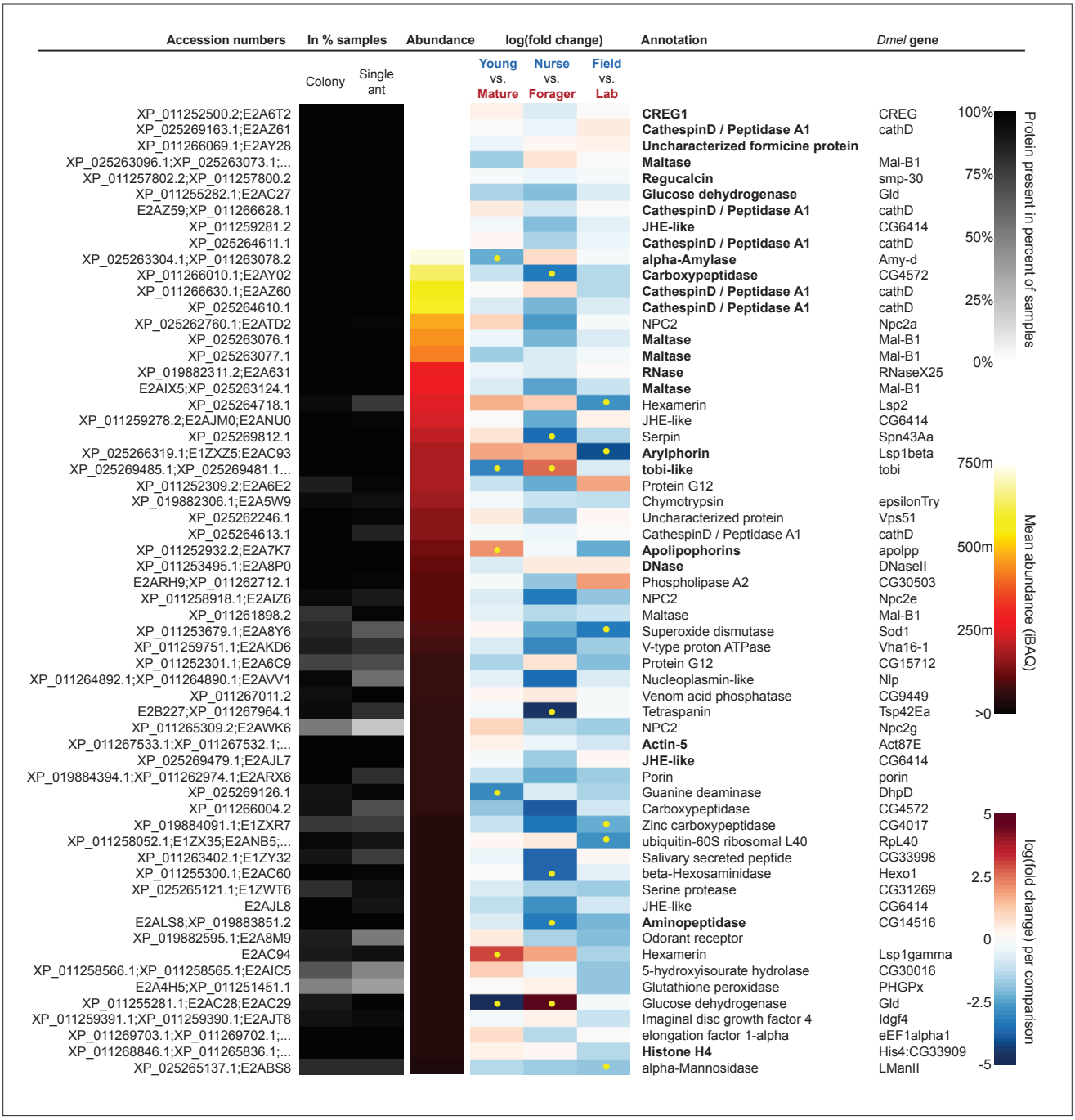

Figure 4—figure supplement 1. Most abundant proteins with accession numbers. The sixty most abundant proteins in trophallactic fluid over 73 colony and 40 single individual samples. Ranking of abundance included zero values. From left to right, accession numbers, proportion of samples in which the protein 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).

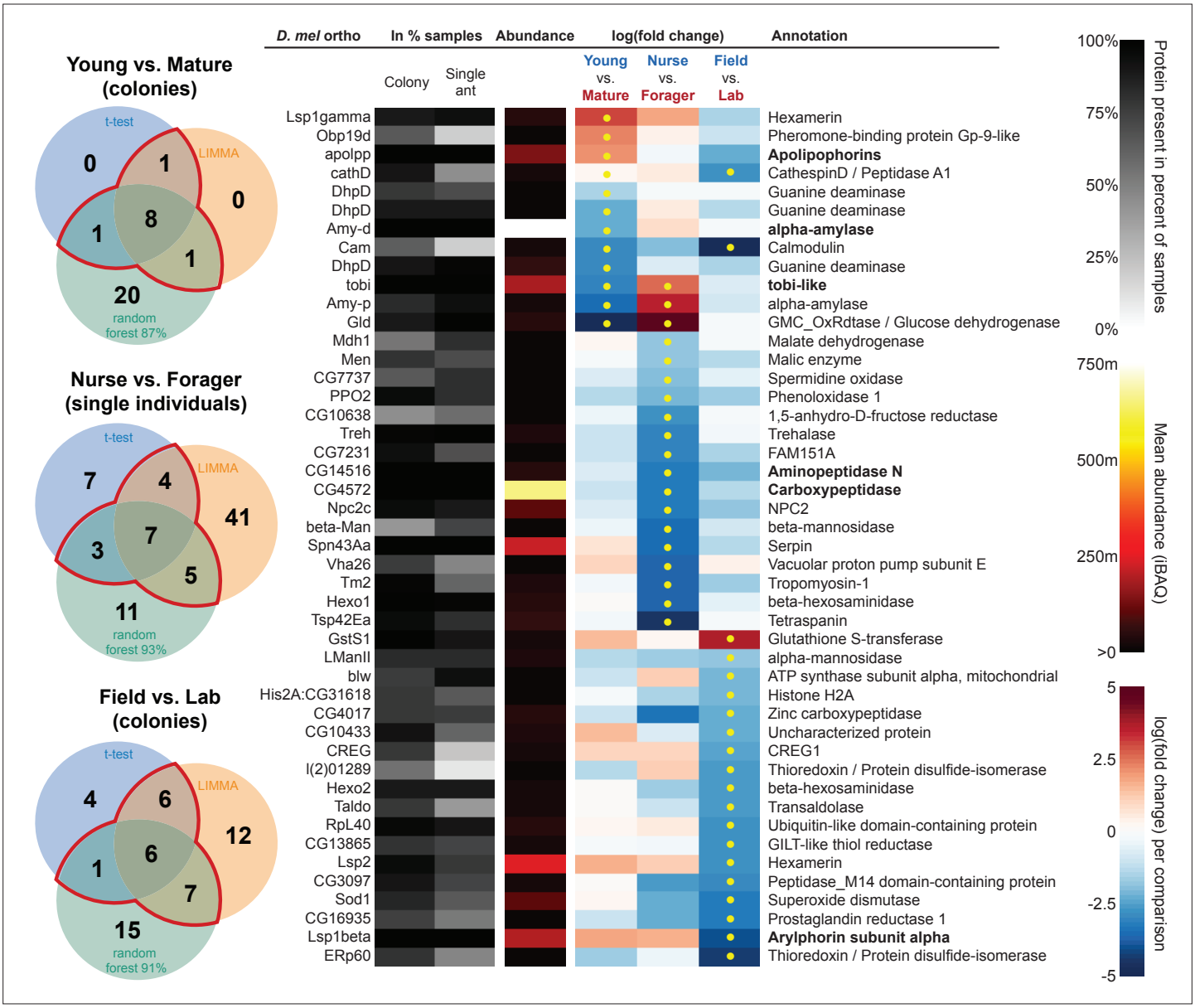

Figure 5. All proteins that significantly differ in two out of three of the analysis methods (frequentist, empirical Bayes, and random forest classification with SHAP values). From left to right, Venn diagrams of significance overlap between methods, *Drosophila melanogaster* orthologs, proportion of 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 ***Supplementary file 2***. For full model results, see ***Supplementary files 3-5***.

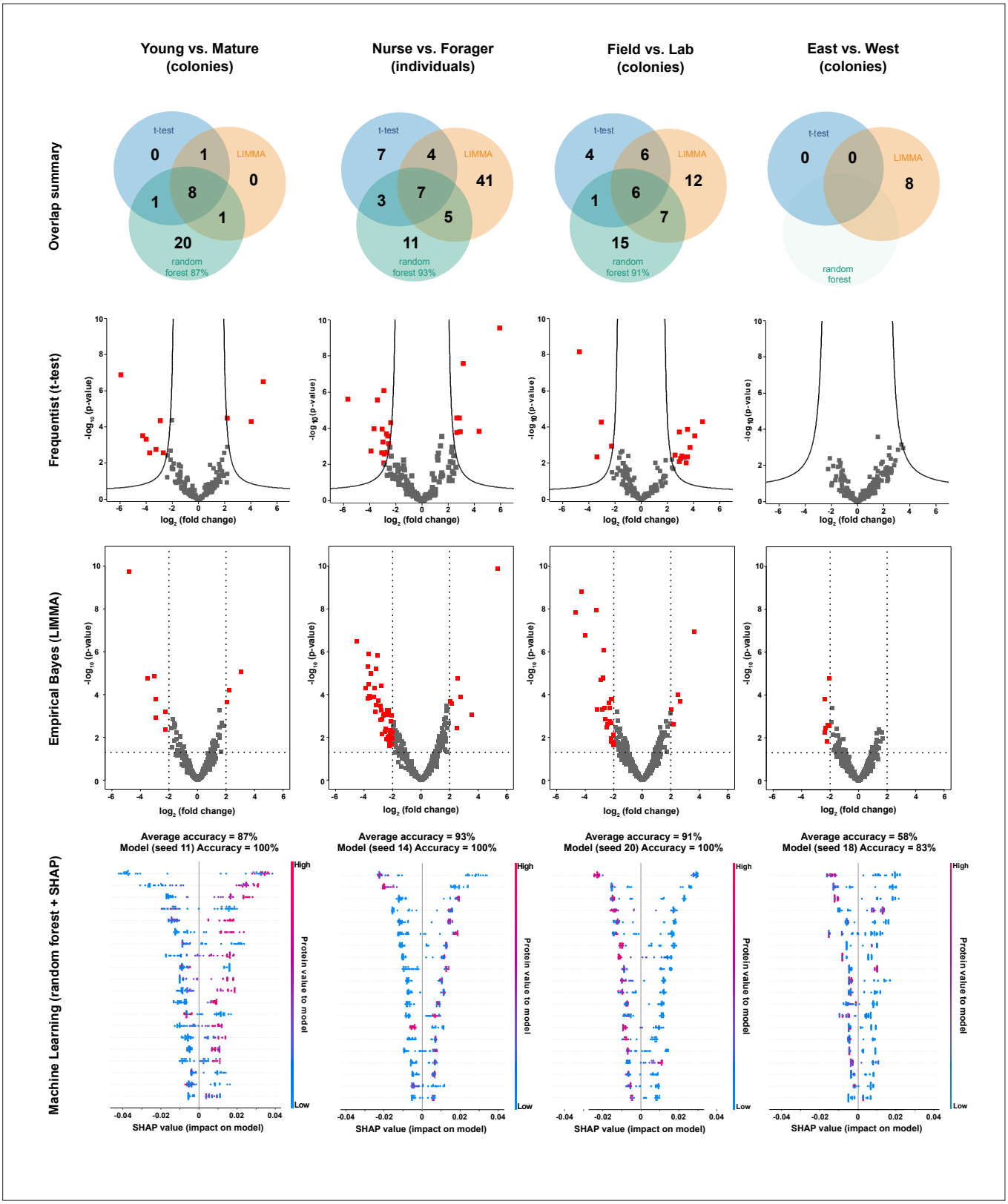

Figure 5—figure supplement 1. Visualization of all results. Venn diagrams summarizing statistical methods, frequentist volcano plots, empirical Bayes volcano plots, example SHAP value plots of feature importance the top 20 proteins. Each SHAP plot is for one of the ten models trained. For significant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

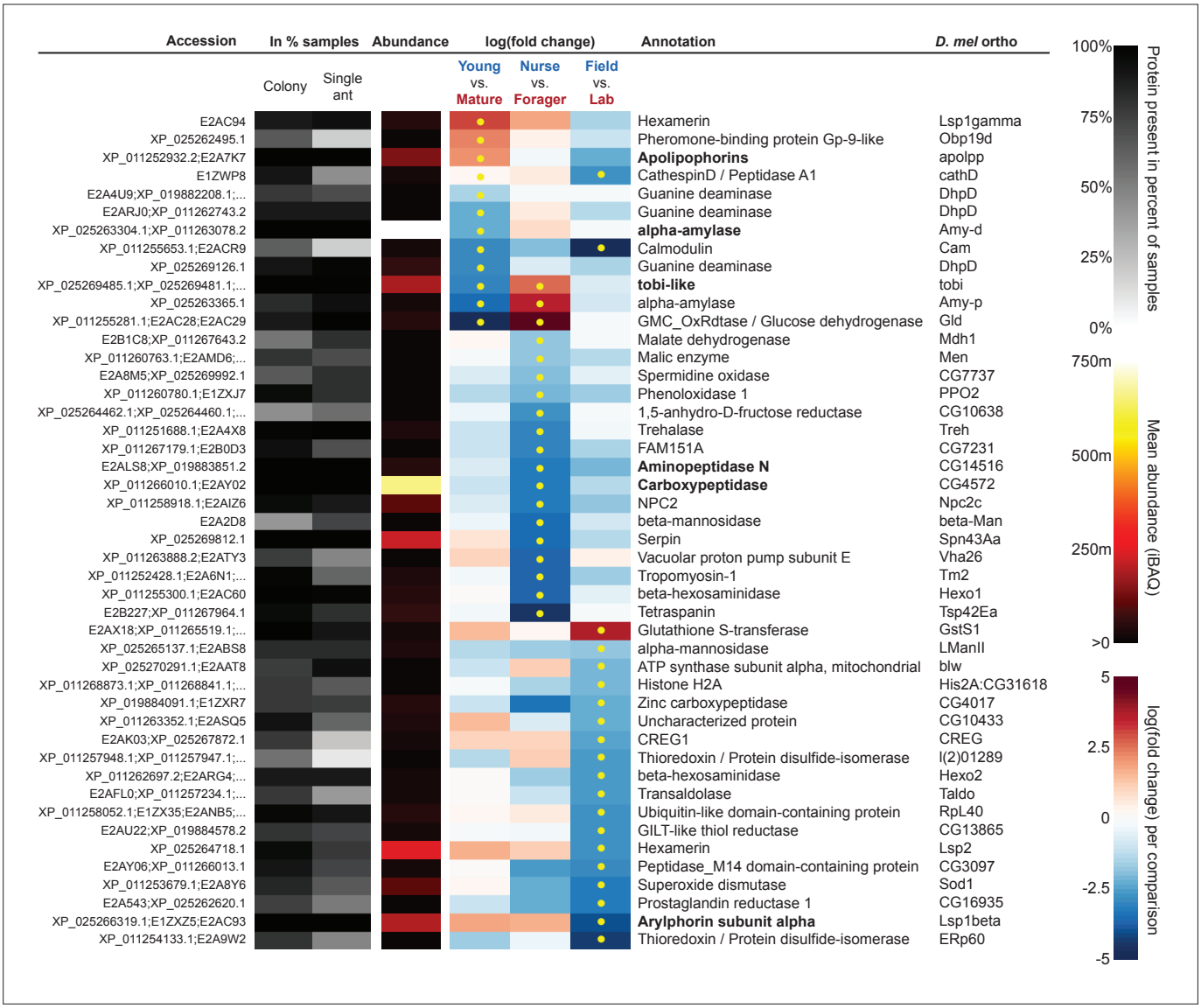

Figure 5—figure supplement 2. Significantly differing proteins in two out of three analyses with accession numbers. All proteins that significantly differ in two out of three of the analysis methods (frequentist, empirical Bayes and random forest classification with SHAP values). From left to right, 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.

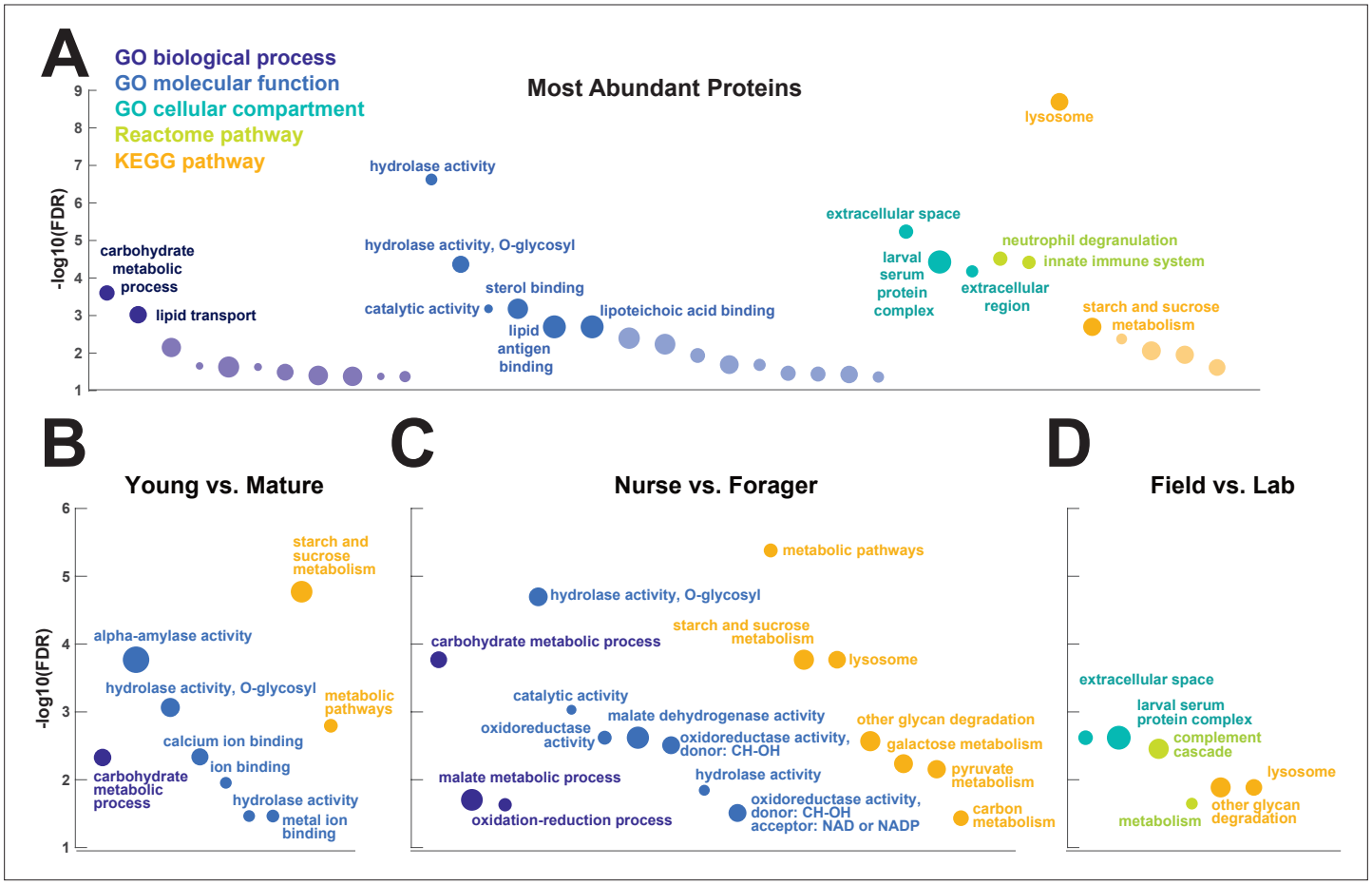

Figure 6. Gene set enrichment analysis of trophallactic fluid. Significant terms for *Drosophila melanogaster* orthologs of (**A**) the 60 most abundant trophallactic fluid proteins, trophallactic fluid proteins significantly differing between (**B**) *Young vs. Mature*, (**C**) *Nurse vs. Forager*, 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—source data 1***.