Caenorhabditis elegans infected with Enterococcus
Chen Chen and Ken Youens-Clark
ABE516 Project 1

Caenorhabditis elegans is a common soil-living nematode which has been shown to prefer eat *Escherichia coli* strain OP50 (Shtonda). This 2018 study compares *C. elegans* fed *E. coli* to groups fed 3 different species of enterococci, “natural commensals of the human gastrointestinal tract and important hospital-borne pathogens, with the majority of human enterococcal infections caused by two species, *Enterococcus faecalis* and *Enterococcus faecium*.” The purpose was to find which *C. elegans* genes were differentially expressed when fed four different bacteria.

The study used several *C. elegans* strains of the wild-type N2 Bristol. There are 13 samples each using 40-50 late L4-staged worms. The study used four groups:

- **C**: Four samples fed heat killed (HK) *E. coli* OP50 (because “live *E. coli* is pathogenic to *C. elegans* on BHI agar, the rich medium required for *E. faecalis* and *E. faecium* growth” [Ausubel])
- **Bs**: Three fed live *B. subtilis* PY79
- **Efs**: Three fed live *E. faecalis* MMH594
- **Efm**: Three fed live *E. faecium* E007

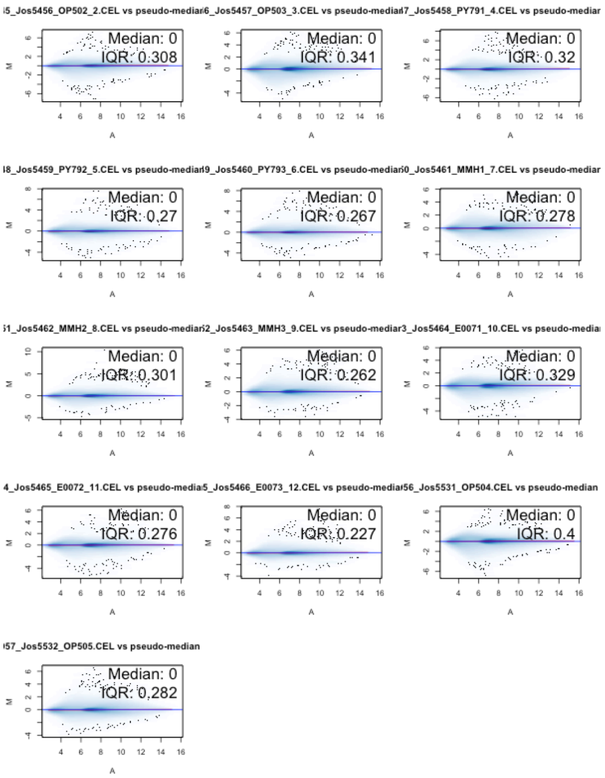
The level of gene expression was determined using the Affymetrix *C. elegans* Genome Array, a DNA microarray platform that measures 22,625 *C. elegans* genes (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL200). The data was obtained from the NCBI Gene Expression Omnibus (GEO) under the accession “GSE95636” (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE95636).

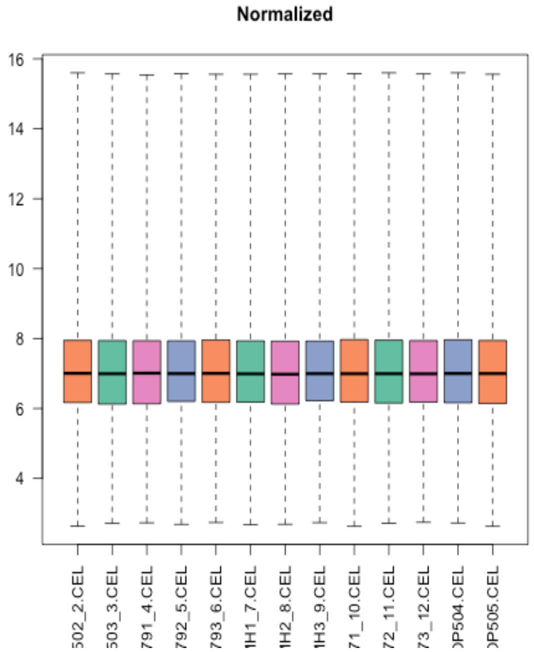
Data Preprocessing

First we checked the quality of the data using density, box, probeset, and MA plots. We found that the distributions of the individual arrays were quite different (e.g., mean, median, standard deviation, and distribution) which makes it impossible to compare measurements from different array hybridizations due to many confounding sources of variation.

To address the problems with data quality, we used background adjustment and normalization. This was essential because part of the measured probe intensity was due to non-specific hybridization and the noise in the optical detection system; therefore, observed intensities need to be adjusted to give accurate measurements of specific hybridization. We used the “RMA” method. We assumed the chips have common distributions, so we used the “quantile” normalization method.

Next, we summarized the Affymetrix platform transcripts by their multiple probes. For each gene, the background-adjusted and normalized intensities need to be combined into one quantity that estimates a number proportional to the amount of RNA transcript. We used the “medianpolish” method. After all these steps, we can take a look at the corrected data. From the MA plot (left) and boxplot (right), we can see that most of the data points are distributed around 0, which satisfies our expectation that most genes are non-differentially expressed. Also, their distributions are quite similar, which makes them comparable.





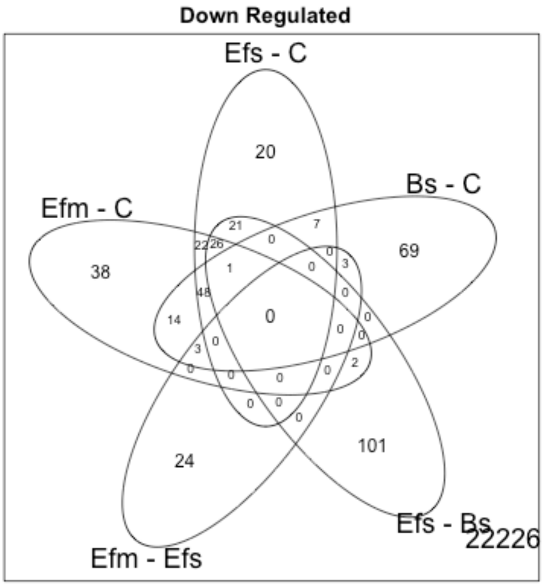
Differentially Expressed Genes

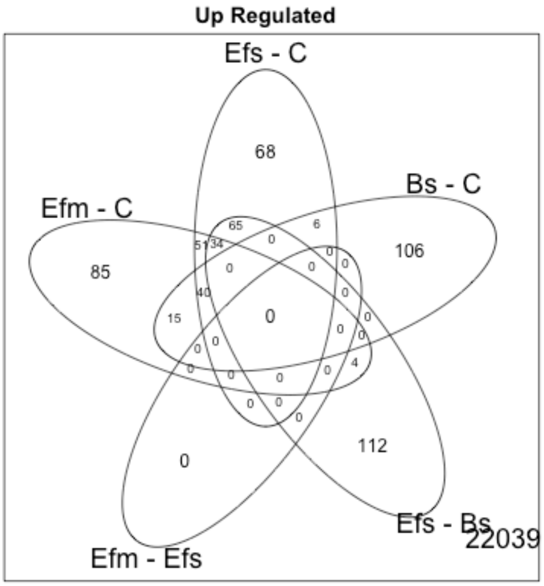
The purpose of a microarray experiment is to find which genes are expressed in a significantly different manner under the various conditions. Using R, it is possible to find the most differentially expressed genes (e.g. top 250) across the entire experiment as well as between each of the contrasts. The tables of these genes in tab-delimited format were written to the “tables” directory.

According to the experimental design in the paper, we constructed five comparisons among four groups. We used a fold-change of 2-fold and a Benjamini Hochberg adjusted p-value of 0.05 as a rejection criterion. Here is a summary table of differentially expressed gene.

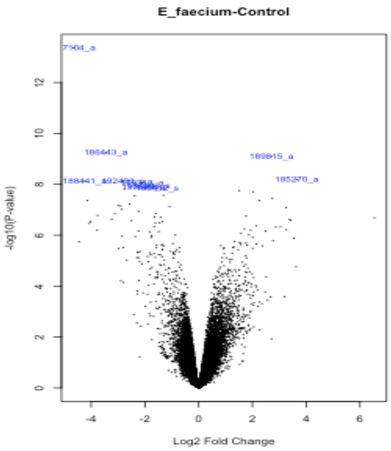
	Efm - C	Efs - C	Bs - C	Efs - Bs	Efm - Efs
Down	154	145	145	151	30
Not Sig	22242	22216	22313	22259	22595
Up	229	264	167	215	0

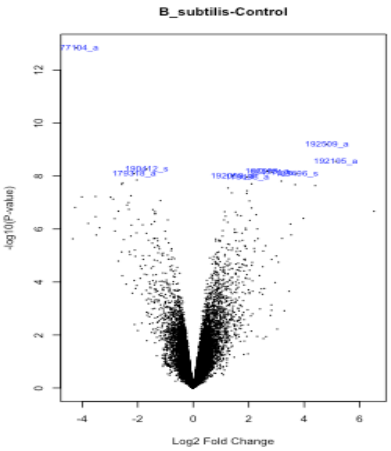
We can use Venn diagrams to inspect the number of down- and up-regulated genes:

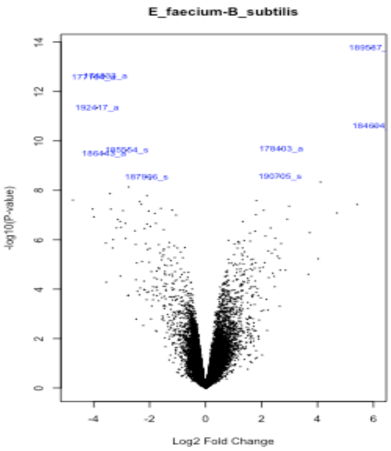


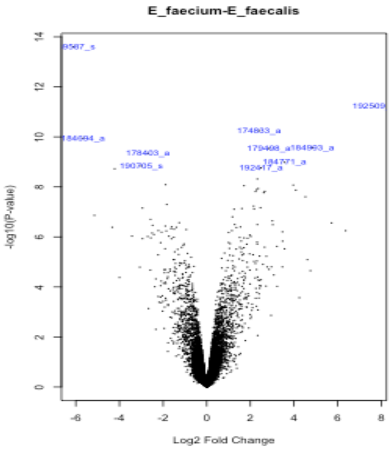


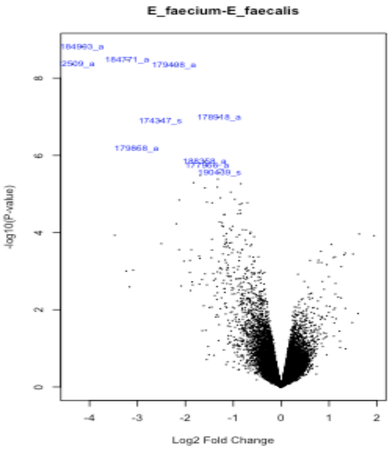
We can also use volcano plots to show highly significant (p-value < 0.05) differentially expressed (FC > 2) genes.





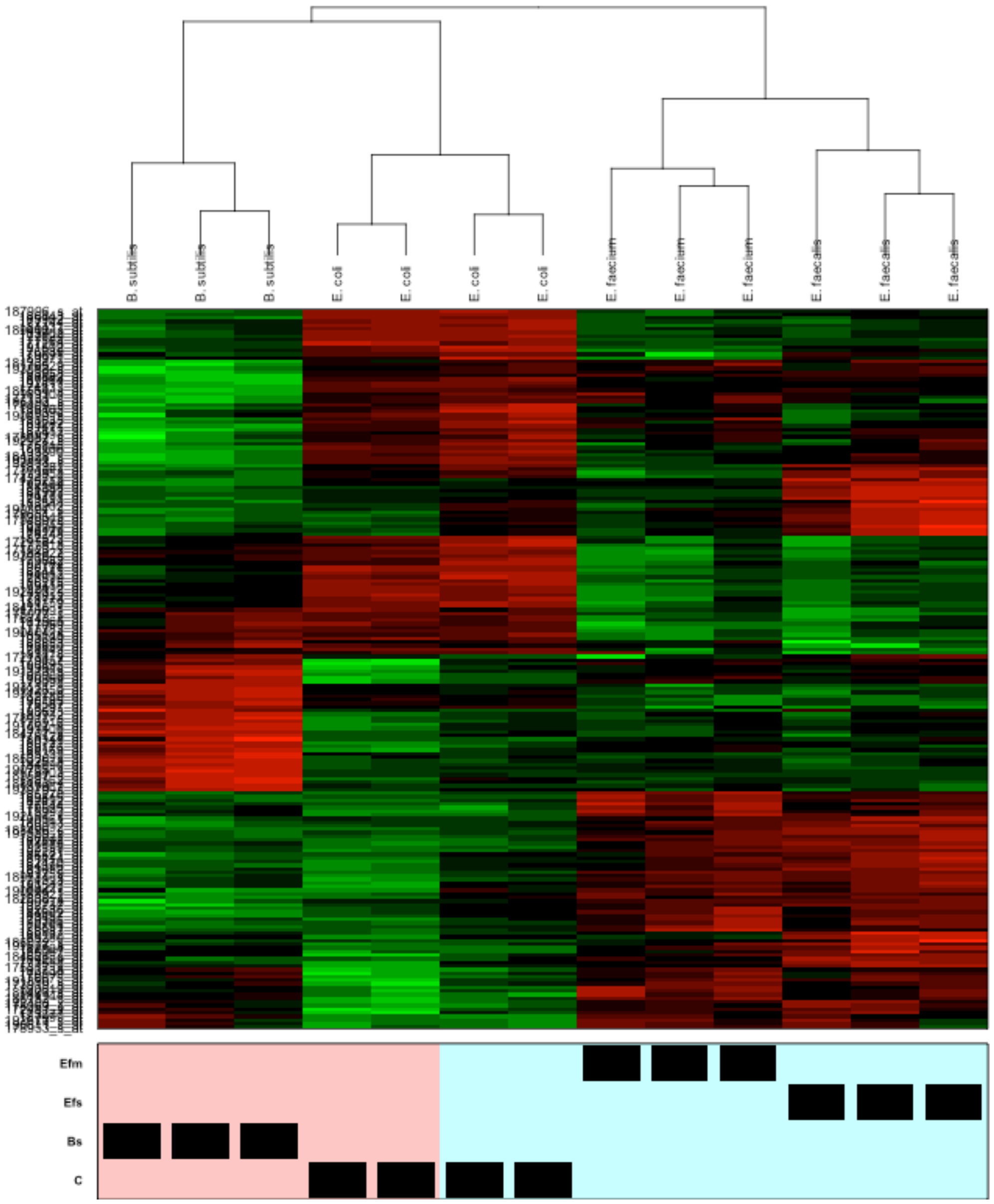






Cluster Analysis

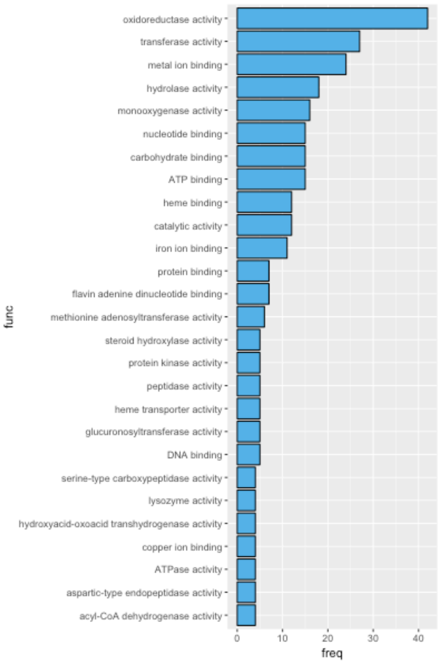
Clustering analysis can identify groups of co-regulated genes by their co-expression patterns. We used the hierarchical clustering method with Pearson correlation distance, the agglomerative algorithm, and average linkage. To find the optimal number of clusters (K), we checked both gap statistics and silhouette methods. They showed the same result, K = 2. Here is a heatmap of final clustering result.

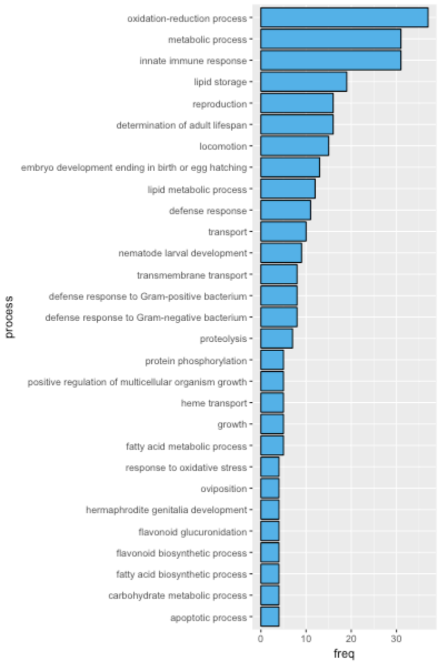


From the heatmap, we can see that for *E. faecalis* and *E. faecium*, their colors follow the same pattern, which means they share many similar genes. Also, from the dendrogram, we see they are grouped in the same cluster. *E. faecium* and *E. coli* (control) are far away from each other in their color patterns. *E. faecalis* and *E. coli* have some overlap. *B. subtilis* and *E.* coli also have some overlap, but their dissimilarity is still very high.

Gene Ontology Analysis

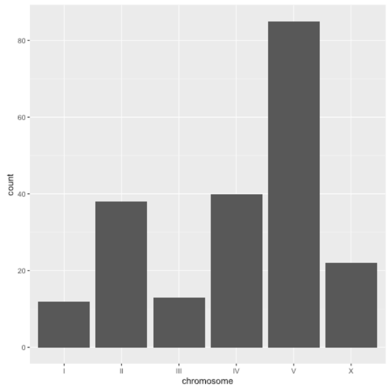
Using the data from the SOFT file, we can get metadata on the genes which includes Gene Ontology annotations. We can summarize the GO processes and functions for the most differentially expressed genes to get an idea of the broad cellular activity being affected.





Chromosome Location

The metadata also provides the genomic location of each feature, so it's possible to see which chromosomes are most affected by the experiment. Using the plot below, it appears chromosome V contains more of the differentially expressed genes by a factor of two.



Conclusion

Using a fold-change of 2 and a BH1995 FDR adjusted p-value of 0.05, we identified 383 differentially expressed genes (229 upregulated genes and 154 down-regulated genes) in the *E. faecium* infection compared with *E. coli*, and 409 differentially expressed genes (264 up-regulated and 145 down-regulated) in the *E. faecalis* infection compared with *E. coli*. The overall differential expression analysis among all 4 groups shows 200 differentially expressed genes. Based on these 200 genes, we conducted a clustering analysis and found some interesting patterns. The clustering of samples is very consistent with the background information of this study. The gene pattern is the focus of this study but needs further biological validation.

References

Ausubel FM, Yuen GJ. Both live and dead Enterococci activate Caenorhabditis elegans host defense via immune and stress pathways. 2018; Virulence, DOI: 10.1080/21505594.2018.1438025.
Shtonda BB, Avery L. Dietary choice behavior in Caenorhabditis elegans. The Journal of experimental biology. 2006;209(Pt 1):89-102. doi:10.1242/jeb.01955.