**NSCI7915: Preliminary Analysis**

**Introduction**

The study by Li et al., *Systematic analyses identify modes of action of ten clinically relevant biocides and antibiotic antagonism in Acinetobacter baumannii*, contains a dataset displaying changes in *A. baumannii* BAL062 gene expression in response to biocide exposure. This dataset possesses from the entire *A. baumannii* BAL062 genome, with gene expression interpreted via the log-fold change (logFC) TraDIS reads for each gene. For ease of interpretation, this genome-wide dataset was split into multiple subsets, with each filtered for genes related to specific function.

This data analysis project focuses specifically on the “Transporters” subset, containing genes related to transporter proteins, as well as their regulators and other associated proteins. This subset of genes contains four unique ATP-binding cassette (ABC) transporters: adeABC, adeIJK, znuABC and ttg2ABC. Each of these transporters consist of three constitutive A, B and C proteins, each encoded by a single gene. By analysing the logFC of these constitutive genes, as well as considering the logFC of the gene families as a whole, this study aimed to investigate the similarities and differences in biocide response between these four transporter complexes.

To characterize such similarities and differences, two main hypotheses were tested. i) that adeABC and adeIJK will have similar biocide profiles, as both are transmembrane RND pumps. Moreover, these biocide profiles will be distinct from znuABC and ttg2ABC (MFS), which are inner-membrane/periplasmic pumps. ii) Due to variations in chemical structure and mode of action, each biocide will interact with each gene family in a unique manner.

The Transporters dataset possesses numerous hypothetical or undescribed proteins. Initial investigative analysis suggested that some of these genes may be associated with some of the transporter families. In an attempt to critique this association, a third hypotheses was tested: iii.) ABC pumps, and their associated genes, have characteristic profiles that can be used as evidence to characterise the hypothetical proteins BAL062\_00181 and BAL062\_01982.

**Methods**

*ANOVA Assumptions*

For each ANOVA conducted the ANOVA assumptions were first tested to assess the validity of the result. To assess the validity of analysis results, the ANOVA assumptions were also tested. Homogeneity of variance was assessed by a Levene’s Test, which tests the null hypothesis that the variance is equal across all groups. This was further tested by plotting residuals against fitted values and assessing the fitted line. The assumption of normal data was tested by assessing the linearity of a Normal Q-Q plot. The third assumption, the independence of observations, is intrinsic to the experimental design. The expression of each gene family is regulated by four independent and unique systems, ensuring that the expression of one is not intrinsically associated with the expression of another. Moreover, each biocide exposure was performed on individual colonies, with the expression of a single gene measured in each colony.

*Hypothesis 1: Differences between ABC gene families*

Principal components were calculated for all genes of the dataset and used to generate a skree plot. Based on this plot, a factor analysis was performed and factor analysis scores were plotted. This plot was used as an initial investigative analysis to determine if genes of the same ABC family are correlated with one another.

To investigate differences in logFC between gene families, a One-way ANOVA was performed. This ANOVA was followed by a Tukey’s test of Honestly Significant Differences (HSD) to characterise the variation between groups. The differences noted by this analysis were visually supported by the comparison of logFC bar plots for each gene family.

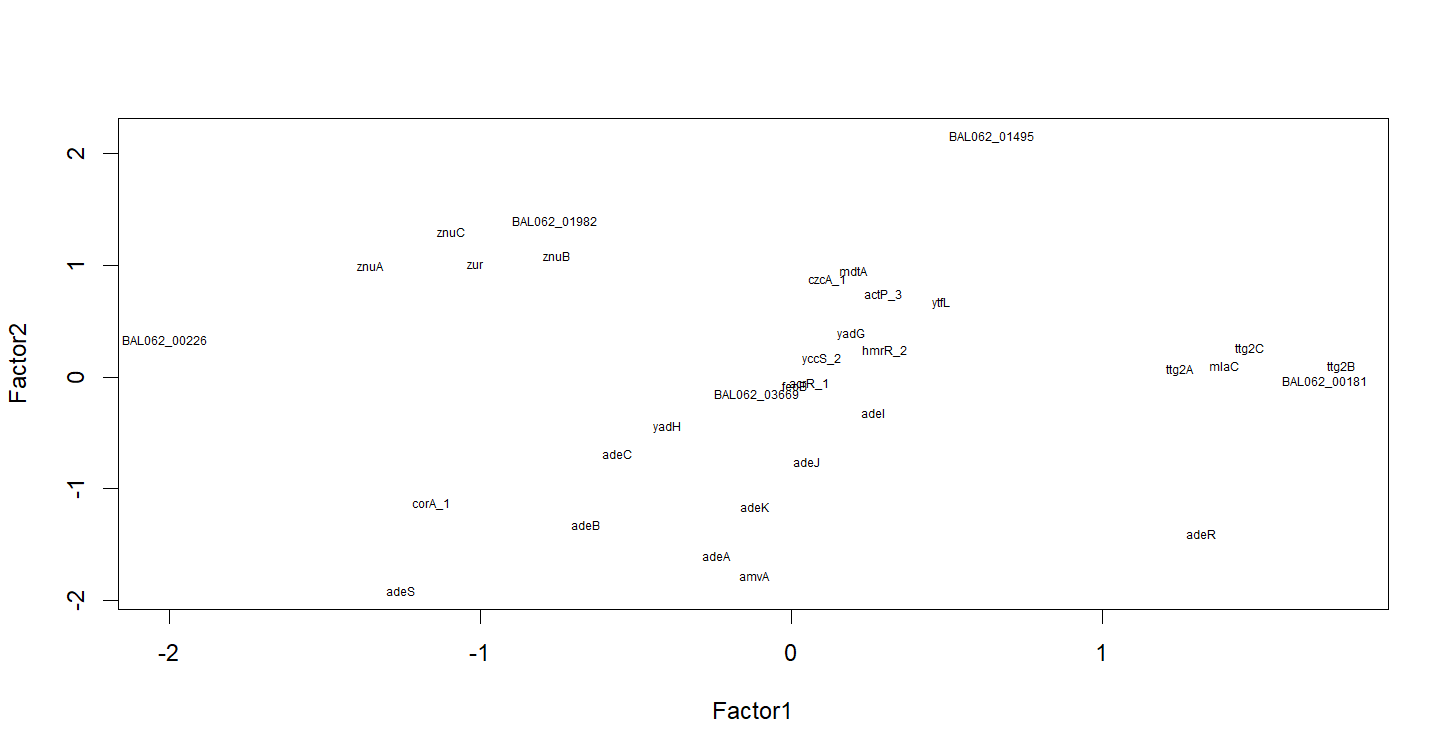
*Hypothesis 2: Characterising biocide effects*

A One-way ANOVA was used to investigate the differences in logFC between biocides. This was further characterised by a Tukey’s HSD. To determine if different biocides interacted with each gene family in a unique manner, a second ANOVA investigated the interaction between biocide and gene family on logFC. These differences were further characterized by a Tukey’s HSD. However, due to the sheer volume of significant interactions, the results of this test were unclear. Instead, a bar plot displaying the logFC for each gene family, grouped by biocide, was constructed were easier visual interpretation.

*Hypothesis 3: Utilising biocide profiles to characterise hypothetical proteins*

The before mentioned investigate factor analysis suggested that hypothetical protein BAL062\_00181 was an associated gene of znuABC, and that BAL062\_01982 was associated with ttg2ABC. This was supported by the fact that the factor analysis plot grouped known associated genes zur and mlcA with znuABC and ttg2ABC genes, respectively. To further investigate this relationship, logFC data for BAL062\_00181, BAL062\_01982, zur and mlcA were included in another One-way ANOVA

investigating the interaction between groups (a stand-in for gene family) and biocide. These interactions were further characterised by a Tukey’s HSD. Finally, the logFC values for each hypothetical protein and the confirmed associated gene were plotted alongside each gene family for a visual assessment of biocide profiles.



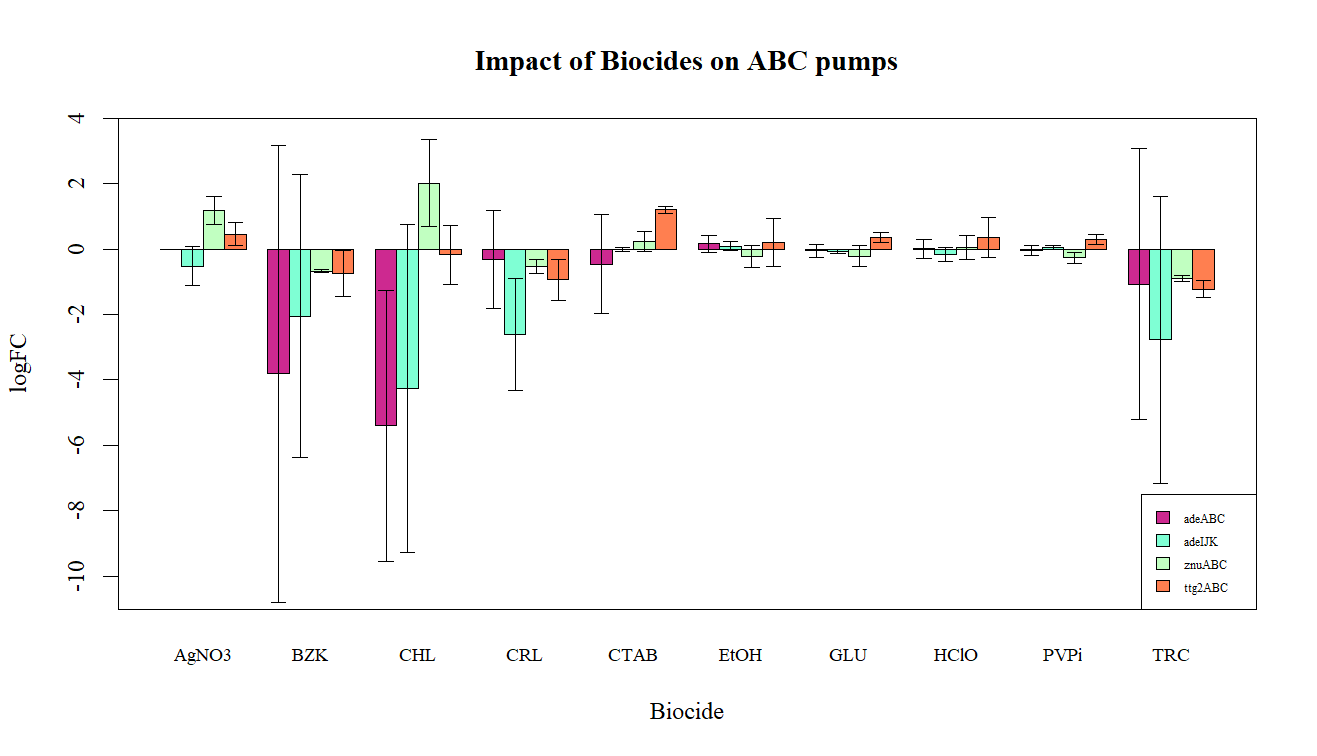
**FIG 1.** Plot of factor analysis scores for all genes of the “Transporters” dataset. Genes from the same family appear the same vague groupings, reflecting how gene function (and consequently, logFC) may be influencing the structure of the data. 

**Results**

*Differences between ABC gene families*

A principal components analysis of the *Transporters* dataset was used to construct a Skree plot (SX). Visual assessment of the plot indicated four factors to be sufficient, and a subsequent factor analysis was performed following these conditions. This analysis did not reject the null hypothesis, that the majority of variation was explained by four factors (χ2(11) = 21.21, p = 0.0313). While this suggests that the fit may require more factors, it is important to consider that each measurement only has one replicate. This low sample number can make the p value unreliable. As such, a plot of factor analysis scores was generated to consider the structure of the data (**Fig 1.**). This plot displays distinct structures and groupings of the data. Most notably, genes of the same family are clustered into unique or near-exclusive groups. This acts as a first line of evidence towards the notion that different ABC gene families possess different logFC profiles.

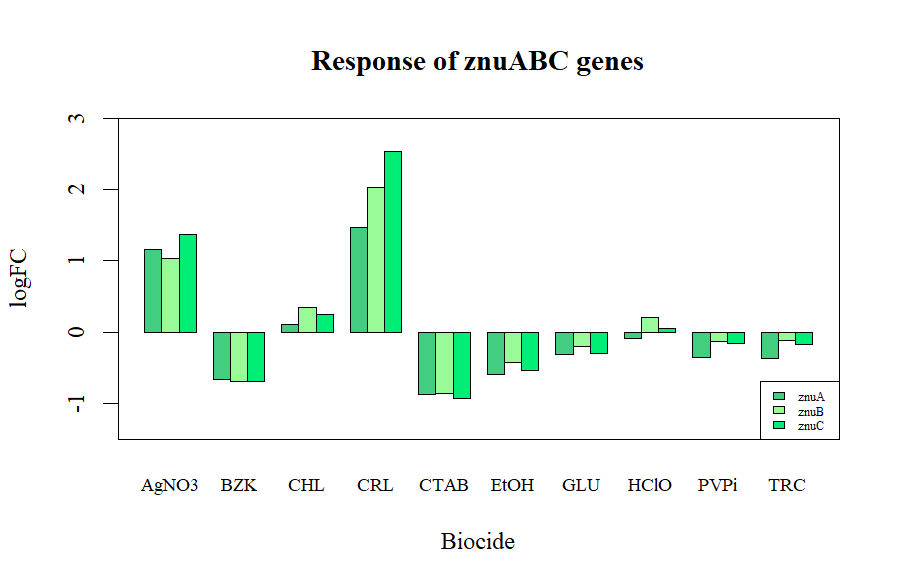
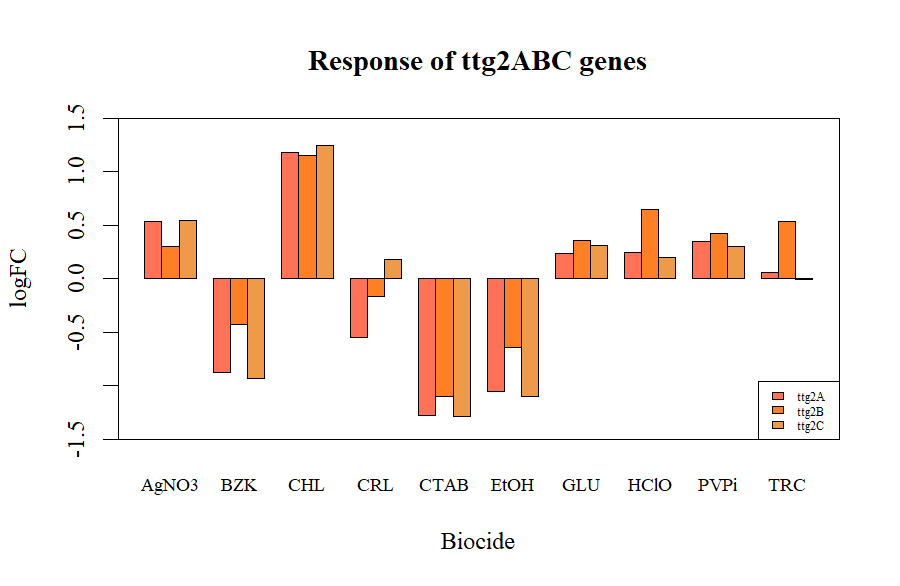
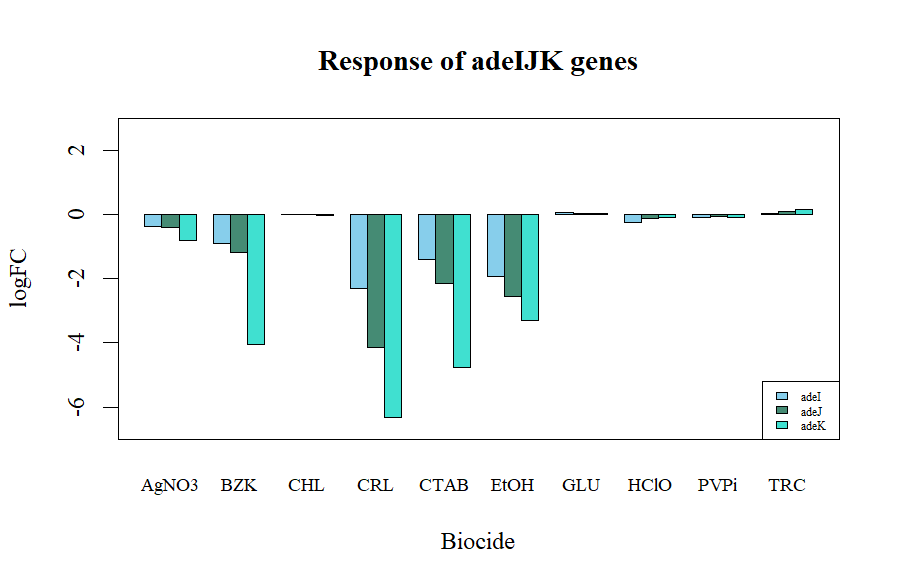
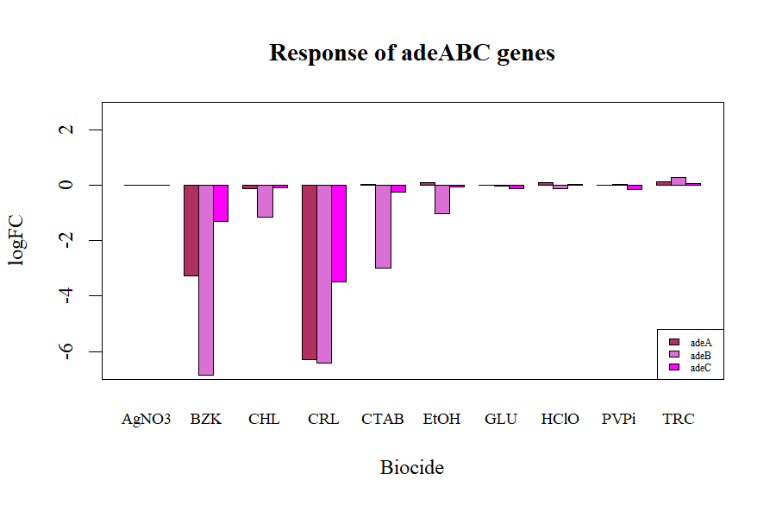
The linear model of gene families against logFC failed the first two ANOVA assumptions. The Residuals vs Fitted Values plot has a slight negative slope, and the majority of data points fell below this curve (SX). Moreover, the Levene’s Test rejected the null hypothesis (F4,145= 2.56, p = 0.04107). Normality of data was also rejected by the Normal Q-Q plot. Linearity was observed within the first quantile, but was lost past this point (SX). However, such failures of the strict ANOVA assumptions are to be expected. As each gene family only contains three data points (one for each A, B and C gene) the data is impossible to be normal. Furthermore, the lack of homogeneous variances can be attributed to the fact that each individual gene of each family has unique logFC values in response to the biocides. However, as meaningful qualitative information can be obtained from the analysis, the ANOVA was still conducted.



**FIG 2.** Mean (±95%CI) ABC gene family logFC measurements, organised by biocide treatment (n=3). Note that the value logFC for each biocide is the average logFC of the three constituent A, B and C genes in each family.

The null hypothesis, that logFC profile does not vary between the four ABC families, was rejected (F3,116 = 6.5688, p = 0.0003854). Tukey’s HSD found the logFC profile ttg2ABC to be 1.08 points greater than that of adeABC (p= 0.027, 95% CI [0.0864, 2.07]), and 1.21 points greater than adeIJK (p=0.010, 95% CI [0.218, 2.21]). znuABC was similarly 1.17 points greater than adeABC (p=0.01, 95% CI [0.177, 2.17]), and 1.30 points greater than IJK (p=0.0048, 95% CI [0.309, 2.30]). However, no significant difference was observed between adeABC and adeIJK, or between znuABC and ttg2ABC. Therefore, the null hypothesis that the profiles of adeABC and adeIJK would not be

distinct from those of znuABC and ttg2ABC, was rejected. The biocide logFC responses of adeABC and adeIJK are more similar to anoe another than to that of znuABC and ttg2ABC. Similarly, the biocide profiles of znuABC and ttg2ABC appear to display some kind of similarities.



**FIG 3.** logFC measurements of A, B and C genes for each ABC gene family. **a)** adeABC, **b)** adeIJK, **c)** ttg2ABC, **d)** znuABC

**a)**

**b)**

**c)**

**d)**

Such similarities and differences are visualised in **Fig 2.** The similarities suggested at in this data appear to be attributed to directionality of logFC (positive or negative change in gene expression) as well as magnitude of logFC change. Another point of difference between the gene families is the variation of logFC within each family, as suggested by the 95% confidence intervals. Intervals for adeABX and adeIJK are strikingly obvious and encompass a wide range, whereas intervals for znuABC and ttg2ABC are more restricted. This indicates a greater variation in logFC of A,B abd C genes within the adeABC and adeIJK families than that of the znuABC and ttg2ABC families. Such variations are displayed in **Fig 3**. While variation exists across all gene families, it is visually apparent that such variation is more explicit in adeABC and IJK than in znuABC and ttg2. However, such variation is also unique between the two ade families. In the case of adeABC, the B gene (adeB) is the more consistent outlier, whereas the C gene (adeK) is the outlier of adeIJK. In contrast, znuABC and ttg2ABC lack such significant outlying genes, further highlighting the distinction between the two grouped of transporters.

*Characterising biocide effect*

Similar to the previously-described model of logFC and gene family, the linear model of logFC and biocide type failed to meet the first two ANOVA assumptions. Variances were non-homogeneous as evidenced Levene Test rejection of the null hypothesis (F9,110= 10.4, p = 1.646×10-11), and a non-linear Residuals vs Fit plot (SX), and the Normal Q-Q plot displays that the data was not normal (SX). But once again, the qualitative nature of the results from an ANOVA would still prove valuable, and so the analysis continued.

The One-way ANOVA rejected the null hypothesis that the logFC effect of all biocides was the same (F9,110= 5.28, p = 5.585×10-6). Therefore, each biocide has a unique logFC, likely reflective of their unique chemical structures and modes of action. A Tukey’s identified significant differences between CHL and AgNO3, CTAB, EtOH, GLU, HClO and PVPi. BZK was also found to significantly differ from this same set of biocides (See SX for descriptive statistics of these significant differences). Such findings are further supported by **Fig 2.**, where BZK and CHL elicit the most visually distinct logFC values.

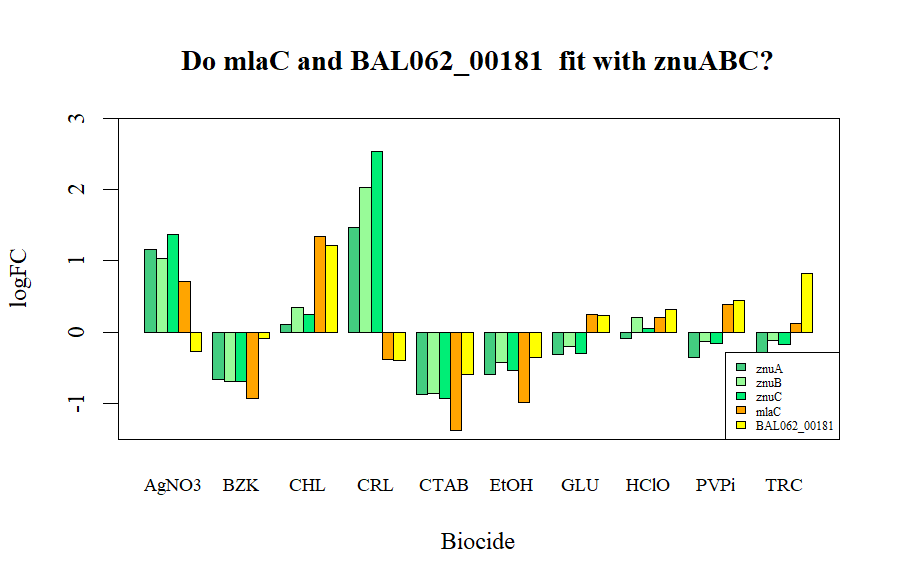
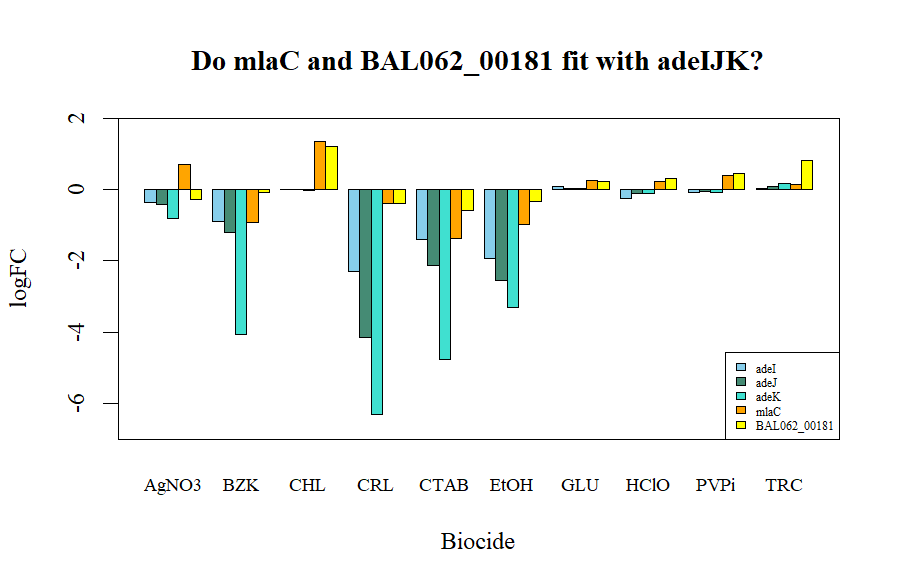
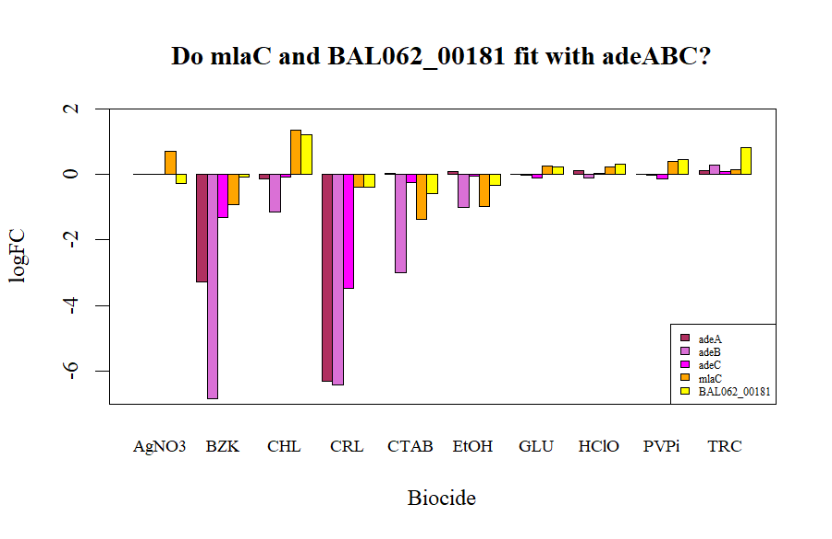
The linear model investigating the interaction between biocide and gene family displayed varied success in relation to the ANOVA assumptions. While the Residuals vs Fit and Normal Q-Q plots (SX, SX) appear poor, the Levene’s Test failed to reject the null hypothesis (F39,80= 5.28, p = 0.0646). Therefore, the analysis of this interaction was performed with slightly more confidence in the quantitative results.

The null hypothesis, that there is no significant interaction between biocides and gene families, was rejected (F27,80= 6.44, p = 3.732×10-11). This suggests that the effect of a biocide on a gene’s logFC is dependent on the gene family. Thus, some ABC transporters are more sensitive to exposure of some biocides over others. A Tukey’s HSD was performed to further characterise these interactions, but the sheer volume of interactions in the output makes interpretation difficult. However, this does not detract from the significance of this result. In fact, such a result just serves to add further validity to the trends and relationships displayed in **Fig 2.** and **Fig 3.**

*Using biocide profiles to characterize hypothetical proteins*

A linear interaction model containing the before-mentioned genes and families, now with the addition of zur, mlcA, BAL062\_00181 and BAL062\_01982, was tested for the first two assumptions. Once again, Residual vs Fit and Normal Q-Q plots display that the model fails to meet the ANOVA assumptions (SX,SX). However, like the previous interaction model, the Leven’s Test failed to reject

the null hypothesis (F7,152= 1.38, p = 0.2162). Thus, the following descriptive statistics can be regarded with an ounce of validity.



**FIG 4.** Comparison biocide logFC profiles of the hypothetical protein BAL062\_00181 and the ttg2ABC-associated protein mlcA against each ABC family: **a)** adeABC, **b)** adeIJK, **c)** ttg2ABC, **d)** znuABC

**a)**

**b)**

**c)**

**d)**

An ANOVA of the interaction between biocide and group (a stand-in for gene family as to not automatically group the new genes with the pre-established families) rejected the null hypothesis that no significant interaction between biocide and gene groups exist (F63,80= 3.70, p = 2.763×10-8). This analysis also once again rejected the null hypothesis that the logFC profiles of the gene groups do not significantly differ from one another (F7,80= 12.6, p = 8.337×10-11). These results serve to confirm that adding these additional genes into the model does not disrupt the previously effects of gene group, and of biocide-gene group interaction, that have been previously established.

A Tukey’s HSD did not identify any significant differences in logFC profile between the new genes and the pre-established gene families. While this helps to confirm that these genes do indeed share logFC profile similarities to the gene families indicated by the factor analysis plot (**Fig 1.**), it does not help to differentiate them from the other gene families for which such associations are unlikely. However, such comparisons are made possible by assessment of the logFC bar plots.