Chart, scatter chart

Description automatically generated

Supplementary 7: no association between the selection pression (median bin pNpS) and the percentage of transposase gene calls in the genome.

Jimmy’s clarification graphs:

Diagram, venn diagram

Description automatically generated

This is somewhat useful, but I need to take the virus out of this graph

Timeline

Description automatically generatedgraph from [the Integron Finder paper](https://academic.oup.com/nar/article/44/10/4539/2516972)

Trashed graphs:

Here is a correlation between biofilm, transposase, and defense mechanism genes. [note: the association between transposases and defense mechanism is not better than random chance: 0.73\*0.5 = 0.365; I will probably only pick the type of transposase that will carry genes. Class II]

Chart

Description automatically generated

Figure 4. The association between biofilm, transposase, and toxin-antitoxin

Chart, scatter chart

Description automatically generated

This page will not be part of the manuscript, just of reference/error checking

Chart, scatter chart

Description automatically generatedChart, radar chart

Description automatically generated

**Good news: number of transposases in a bin and the abundance of each transposase is independent.** Say a bin has 1 transposase, and its abundance is 20 reads per base pair. Another bin has 40 transposases, the average coverage of those 20 transposases will also be 20 reads per base pair.

Graphical user interface, application

Description automatically generated

Chart, scatter chart

Description automatically generatedSame graph, but no bin names. A dot is a bin

Malaspina crew samples deep ocean only, and for each site (same depth), they sequence samples from a 0.2-0.8μm and 0.8-20μm filter. The bins from 0.2-20 μm filter have a higher count of transposases than those from 0.2-0.8μm (supplementary 5).

Pn-ps ratio: I found that transposases have significantly higher pn-ps than their bin’s median. Besides that, no significant result has been found.

HGTector: shows a random scatter between count of transposase in bin and number of Horizontally Transferred Genes. Need to know whether it’s the case with RNA expression.

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | Non-integron | Defense Mechanism on integron | Other cassette with gene calls |
| Median pNpS | 0.104 (n=300K) | 0.0829 (n=55) | 0.143 (n=463) |
| 2-sample t-tests p-value for difference | | 1.0e-06 | |

Both filtered out pNpS > 10

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | Non-transposase-or-cassette | Transposases | Cassette Seq |
| Median pNpS | 0.105 (n=100K, random-sampled) | 0.176 (n=1121) | 0.194 (n=1236) |
| 2-sample t-tests p-value for difference | | < 1e-30 | < 1e-30 |

Chart, box and whisker chart

Description automatically generated

Figure 5 (excuse the lettering). (D) The proportion of each category of genes in each Malaspina’s bin, separated by their lifestyles. A Free-living bin = its abundance is 10 times higher in 0.2-0.8μm than in 0.8-20μm samples. A particle-attached bin = its abundance is 2 times higher in 0.8-20μm samples. A mix bin = 0.8-20μm abundance between 0.5 to 10 folder compared to the 0.2-0.8μm samples

In deep Malaspina’s samples (bins are all from deep ocean) particle-attached bins are more enriched in transposases and defense mechanisms. Particle-attached microbes and mixed, both have significantly higher proportion of defense mechanisms genes compared to free-living microbes.

In our analysis of the deep Malaspina ocean, defense mechanisms genes are found more in particle-attached MAGs (compared to free-living bins, p-value = 0.0003, 50 bins in both groups, see supplementary 5). [Note that in the literature: toxin-antitoxin system sometimes upregulates or downregulates Biofilm expression, it really depends on the toxin. So a toxin to biofilm relation needs more scrutiny]. They can potentially help their host to compete on a particle.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

The *pN/pS* of cassette genes were significantly higher than those of the normal ORFs

Both transposases and cassette genes have significantly higher *pN/pS* values compared to non-transposase, non-cassette genes. The median *pN/pS* for normal ORFs is 0.105, 0.176 for transposases, and 0.196 for cassette sequences. Out of 1957 cassette sequences with COG functional category calls, only 528 of them have a ≥ 20 coverage for *pN/pS* calculation. From this shrunk sample (we started with 7294 cassette genes), genes belonging to the  “defense mechanism” have a significantly lower *pN/pS* ratio compared to genes in other COG functional categories. The cassette median *pN/pS* for defense mechanisms is 0.0829, 0.143 for other cassette sequences with COG functional calls.

Chart

Description automatically generated

Figure 7: A. *pN/pS* ratio of normal, transposase, and cassette genes. B. the *pN/pS* among cassette genes with identified COG categories: comparing defense mechanism genes to genes belonging to other functional categories. For both graphs, we excluded any *pN/pS* above 10. Observe that cassette genes with COG calls have significantly lower *pN/pS*. Thus, the defense mechanism is only compared to other cassette genes with COG calls  (as opposed to all 1276 cassette genes with pNpS).

Figure 6. The expression rate of biofilm and defense mechanism genes, separated by depth. The expression rate is calculated by log(Normalized RNA Abundance / DNA). For both categories of genes, the expression rate in the mesopelagic zone is significantly higher than these in surface and deep chlorophyll maximum.

The expression rate of biofilms and defense mechanisms are higher in the mesopelagic zone.

**Chart, box and whisker chart

Description automatically generated**

Supplementary 5: particle-attached bins are more enriched in transposases and defense mechanisms genes in malaspina bins. Particle-attached microbes and mixed, both have significantly higher proportion of defense mechanisms genes compared to free-living microbes.

A Free-living bin = its abundance is 10 times higher in 0.2-0.8μm than in 0.8-20μm samples

A particle-attached bin = its abundance is 2 times higher in 0.8-20μm samples

A mix bin = 0.8-20μm abundance between 0.5 to 10 folder compared to the 0.2-0.8μm samples