**Preliminary result of objective 1**

**This file shows 5 parts of analysis:**

1. **sequencing data bioinformatics:**

Table1: Filtering steps and data information

1. **negative control samples**

Fig.1: Taxon composition in control samples, A is read counts and B is read proportion.

1. **sequencing data of tank**

Fig.2: **Taxon composition of samples, A. presents the read proportion of each species in each sample, and B. presents the read counts.**

**4. timeline experiment and detection efficiency**

Fig. 3:Comparing the efficiency of species detection in each pool.

Fig. 4:Bubble plot comparing the efficiency of species detection in each pool.

Fig. 5:Comparison of detection rate among water sample and three sponge species.

**5. Reef experiment**

Fig6: Reef experiment.

Sponge samples don’t have positive detections, so I will ignore reef experiment.

**Explanation:**

Unassigned OTU: OTUs which assigned below Order (no family, genus and species info.) are tagged as unassigned.

Read count: raw read numbers.

Read proportion: OTU reads**/**total reads in this sample, Sum read proportion in one sample=1 or 0(sample without any no OTUs)

Incidence: read count converts to 0/1 data.

target species: species put onto tanks.

Timeline experiment analysis: take only the target species for analysis.

**Summary:**

1. The detection efficiency of water sample decreases significantly over time, and followed by cream sponge (Figure 5A). The detection efficiency of pink sponge was not affected by time in a short time period. The read count of DNA in pink sponge is more like a wave drop than a linear drop (Figure 5B). There is no difference in the overall detection efficiency among the three of them (water sample, pink sponge, and cream sponge, Figure5C). Black sponges don’t have positive detections (Figure 3).

The sponge may act as an accumulator to continuously collect DNA, so species are sometimes detected later than water samples, and the DNA collected by the sponge does not decrease significantly in a short time period (Figure 4).

2. Control problem:

2.1. There is a tank A species show in tank B and C (Zebrasoma), and obviously it has high reads. This could be a real contamination or whatever reason this species or its DNA hides in tanks (Figure1A and 3).

2.2 There are some species detected in tank experiment are no in species list, some of them are reef experiment species, some are not in both lists, this may because the reference database problem or tag-jump or real contamination (Figure 2A).

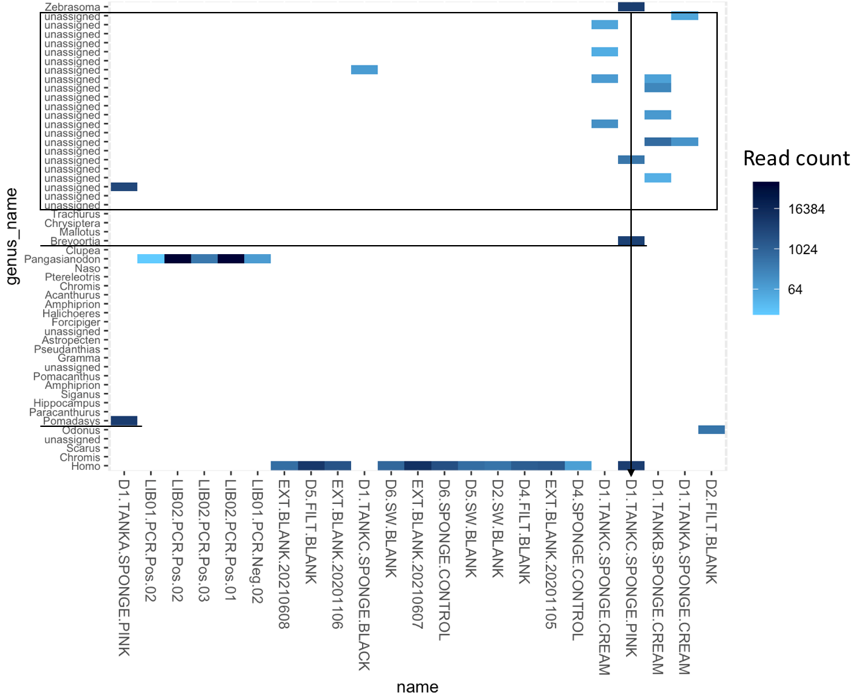
**Table 1: Filtering steps and data information**. A total of 135 samples were sequenced, and 99 got sequence before any clean-up steps. After several clean-up steps, the OTU table contains 78 samples (at least contains one OTU) and 52 OTUs (at least contains 50 reads) for the subsequent analysis. Unassigned indicates the OTU is not assigned to any taxonomic category below Order. Hereafter call them unassigned OTU.

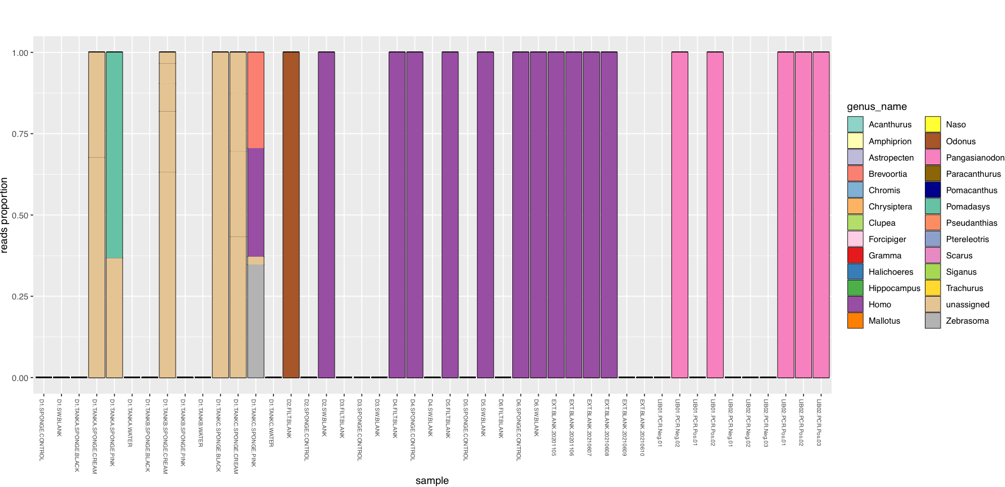
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Raw data** | **Length trim** | **Cluster**  **swram** | **Ecotag**  **assignment** | **Collapsed by**  **LULU** | **Collapsed by**  **metabarpark** | **cleaned by decotam** | **cleaned by phyloseq** | **Final** |
| NO. samples | - | 99 | 99 | 99 | 99 | 99 | 99 | 78 | **78**  **Sp:35 W:15**  **C:23 Reef:5** |
| Total sequences | **1,828,835**  351908+  1476927 | 739,276 | 739,276 | 725,428 | 725,428 | 725,428 | 690,373 | 688,751 | **688,751** |
| NO. OTU | - | - | - | 209 | 131 | 120 | 110 | 52 | **52** |
| unassigned | - | - | - | 116 | 91 | 91 | 81 | 24 | **24** |
| parameter | - | 130 ~190 | -d 3 | Default | Default(0.84) | Default(0.5) | Default(0.5) frequency | lib3>10  lib4>20  otu>50 |  |

\*discarded OTUs are unassigned OTUs. *Gramma\_loreto* is corrected manually, two *Zebrasoma* OTUs are collapsed manually.

**Figure1: Tax composition in control samples, A. is read count and B. is read proportion**. **A.** Y-axis represents Taxon (Genus), the X-axis represents sample-name, colors represent read-count in each OTU; **B.** Y-axis represents the proportion of each Genus in each sample, the X-axis represents sample-name, color codes Taxon (Genus). There is only one OTU in each Genus, so that I will use Genus as taxon index hereafter. Day1 tank control samples have some unassigned OTUs, and two species are not in the experiment list and one in the list. This indicates the initial tank environments are not DNA-free.

The Zebrasom (a tankA species) in the tankC-pink-sponge-control shows positive (in tankB-pink and tankC-cream also show positive, see below), and high reads accompanied these events. It could be actual contamination or whatever reason this fish or its DNA hide in the tank or sponge itself. Fortunately, although Zebrasom is a strong noise, it only shows stochastic at the begging of the experiment, so I will ignore it temporarily in the preliminary result. Blank controls have high human DNA, but it is ok. Usually, I don’t use it as a filter threshold, but human DNA will grab some sequencing space. One PCR negative blank has PCR positive species and even contains higher reads than the PCR positive control. This is probably a TRUE contamination and needs to be discussed how to report it, and this also highlights the lab work procedure needs to reconsider.

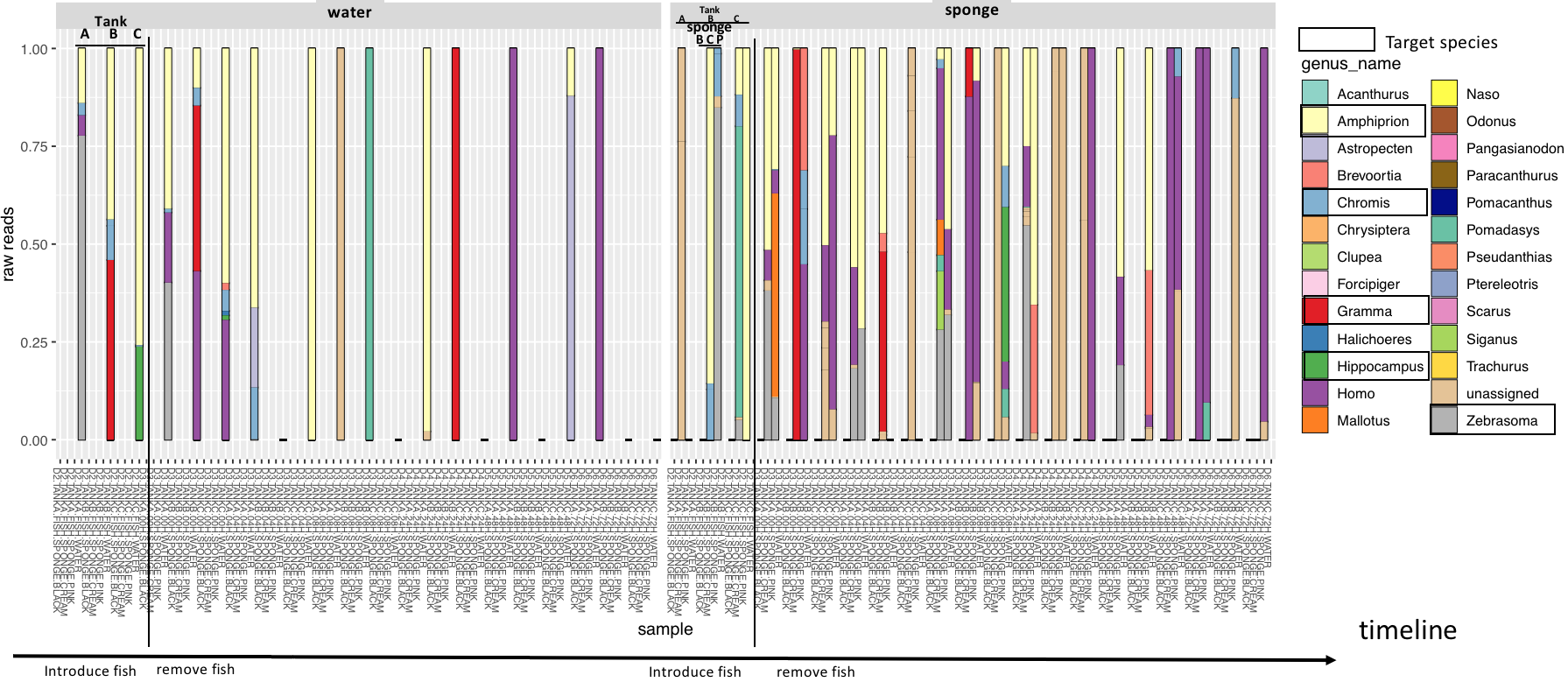
A

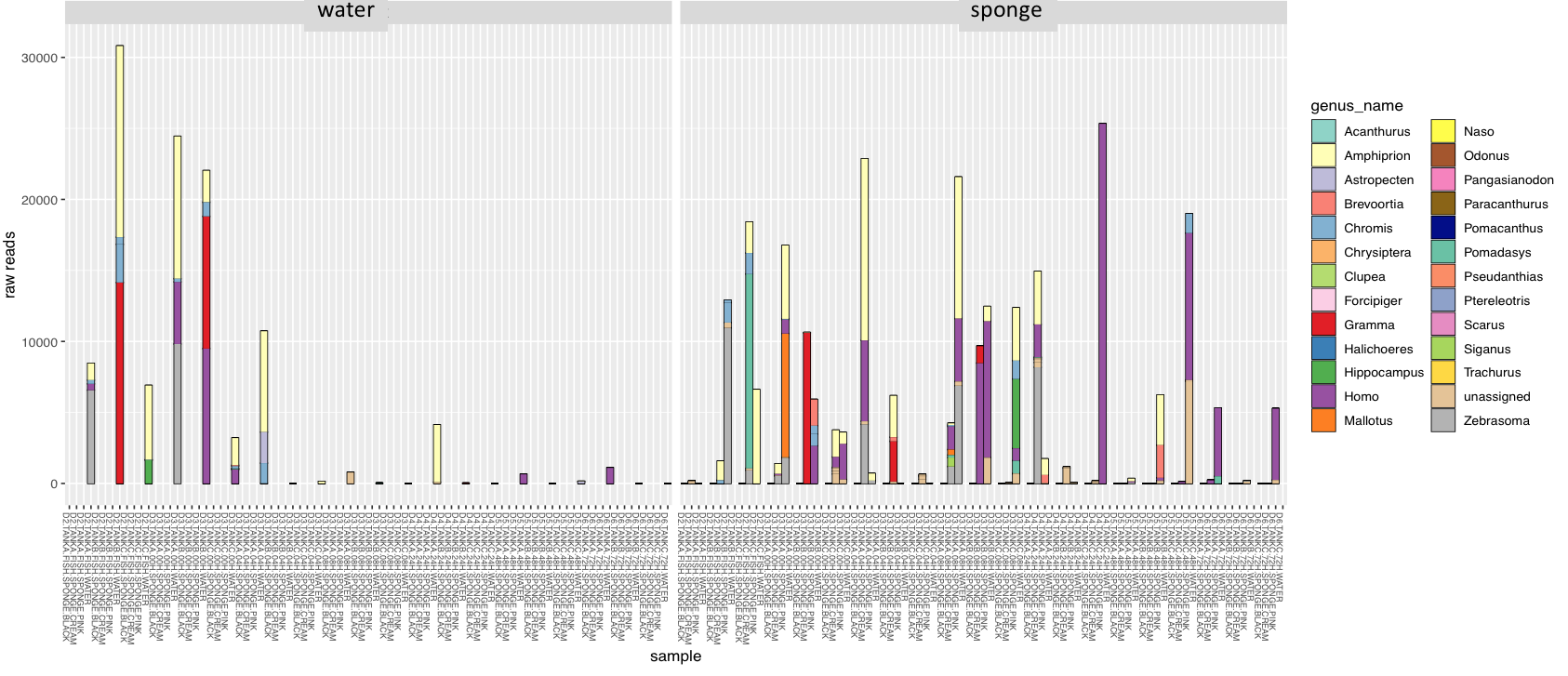
B

**Figure 2. Taxon composition of samples. A. presents the read proportion of each species in each sample, and B. presents the read counts.** The samples are arranged chronologically from left to right. The Black dotted line indicates whether there are fish in the tank (the left side is yes, the right side is no). The left panel is water samples cluster by tanks, and the right panel is sponge samples cluster by sponge type and tank.

Overall taxon composition of water samples is lower than sponge samples, and the detection signal of all target species is clear in the water samples when introducing fish. The total read-count and read-count for each species decreased significantly over time in water samples. The sponge samples didn’t detect all the fish immediately; instead, the detection is lagging. Meanwhile, after removing the fish, the taxon composition becomes complicated. It may cause by two reasons: the ecotag assigns sequences to the wrong species; the other reason is sponge digestion “create” these sequences, and they are not because of contamination and bioinformatic errors.

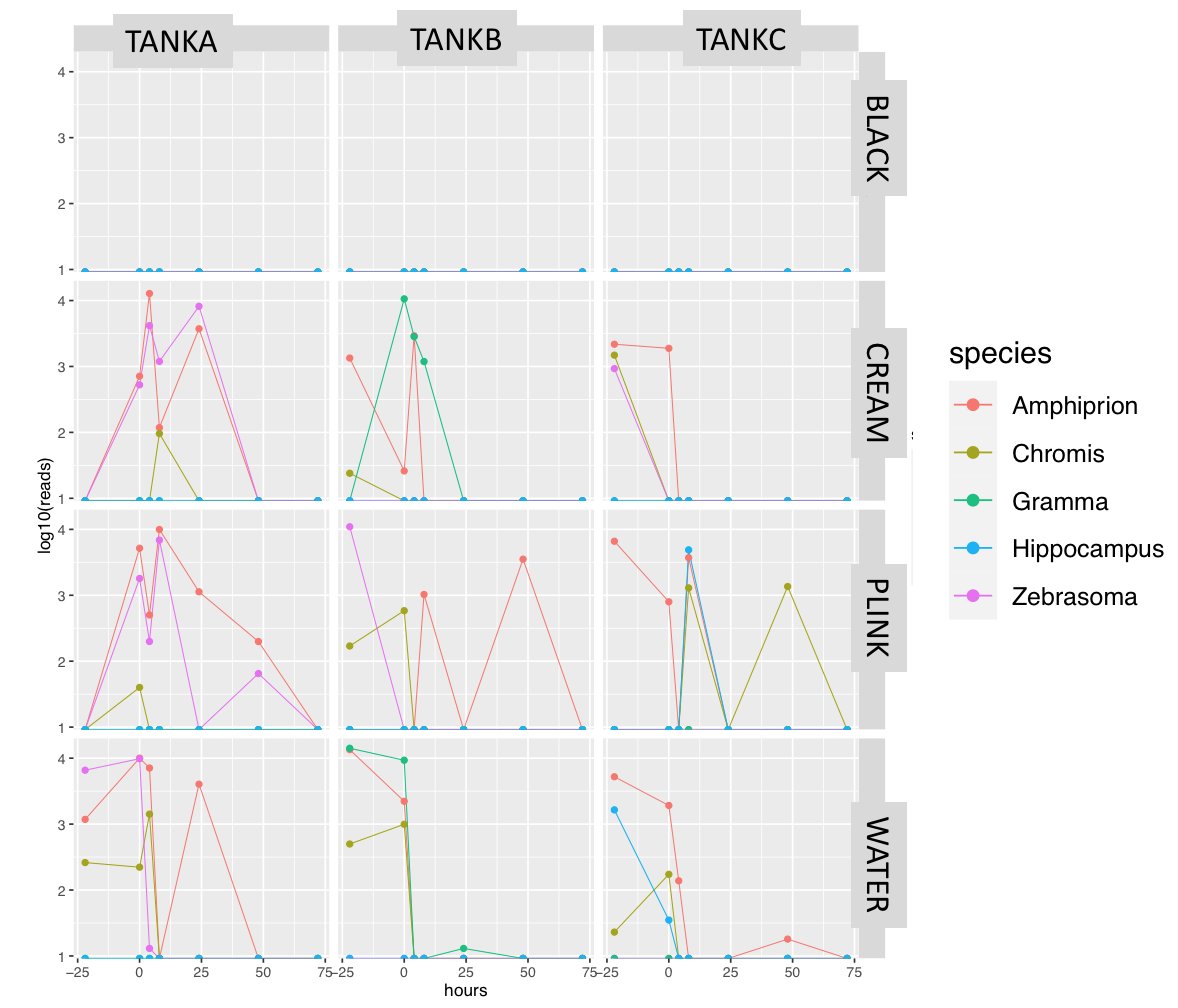
A

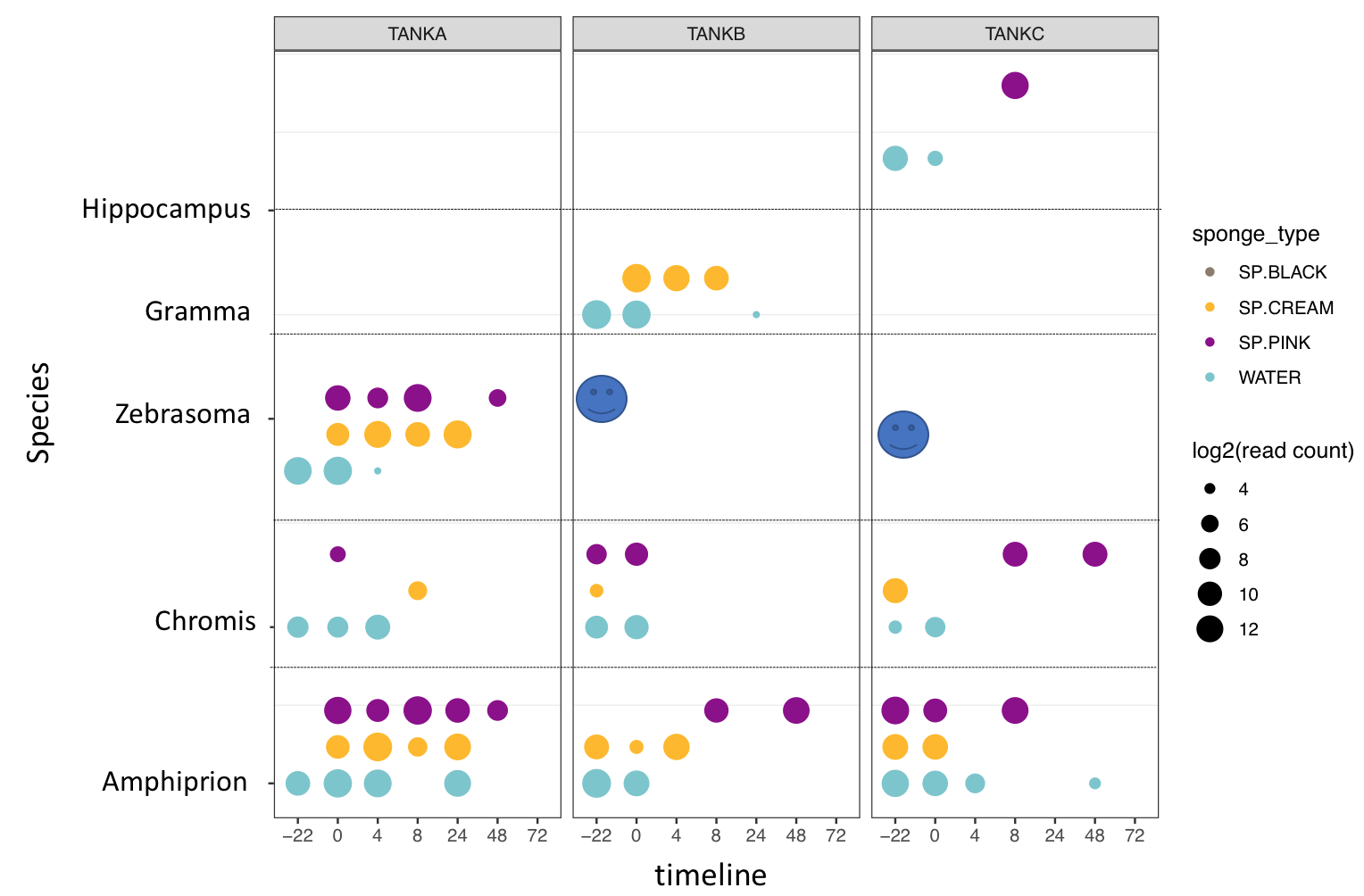


B

**Figure 3: Comparing the efficiency of species detection in each pool.** Figure shows**Read counts of each species in each sample in the timeline experiment.**Column panel represents tank, and row panel represents sample type. Y-axis is the log10(read counts), and X-axis arranges by time elapse. Only the first point indicates fish in the tank, and the rest are without fish. Only experimental list species are used in the figure, other species not in the list but detected are removed from the data. TankA species are Amphiprion, Chromis, and Zebrasoma; TankB species are Amphiprion, Chromis, and Gramma; TankC species are Amphiprion, Chromis, and Hippocampus.

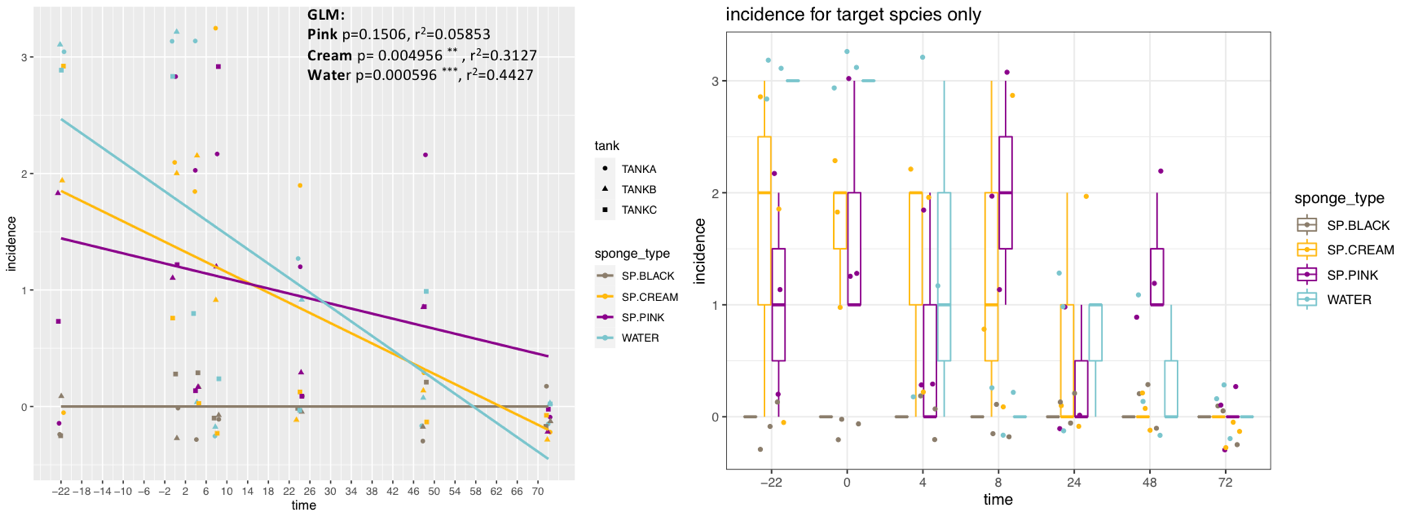
No species in black sponge was found. Both water and sponge-pink & cream successfully caught three species in thank A. The species signal in the sponge is longer than the water sample. Water samples show good detection efficiency in the beginning, and then DNA degrades fast. The pink sponge shows a bimodal pattern, but this isn’t a quantitative experiment (ex. qPCR), so the DNA detection pattern remains to be determined. For the Cream sponge, the pattern of DNA detection is less obvious.

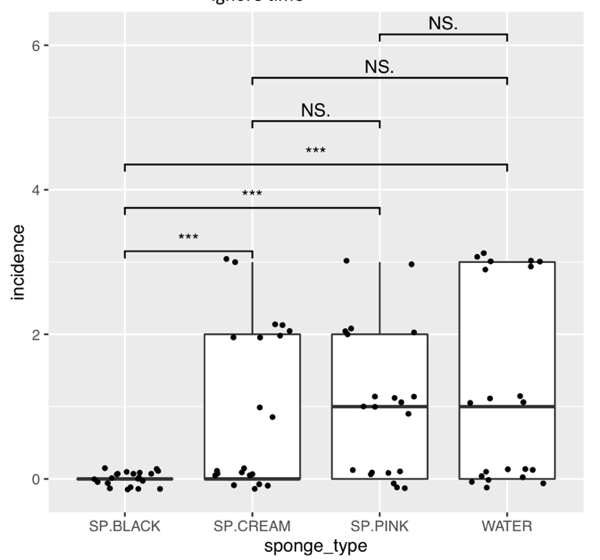
Because the probability of species detection in water samples is not strictly continuous, I am not sure what caused the uncontinuous positive detection in both water and sponge samples. It can be a lab-work effort (extraction and PCR replications, stochastic capture of DNA) or the mechanism of sponge filtration and DNA degradation in sponge cells. It is interesting to discuss the experiment treatments and put the DNA concentration data together on following analysis.

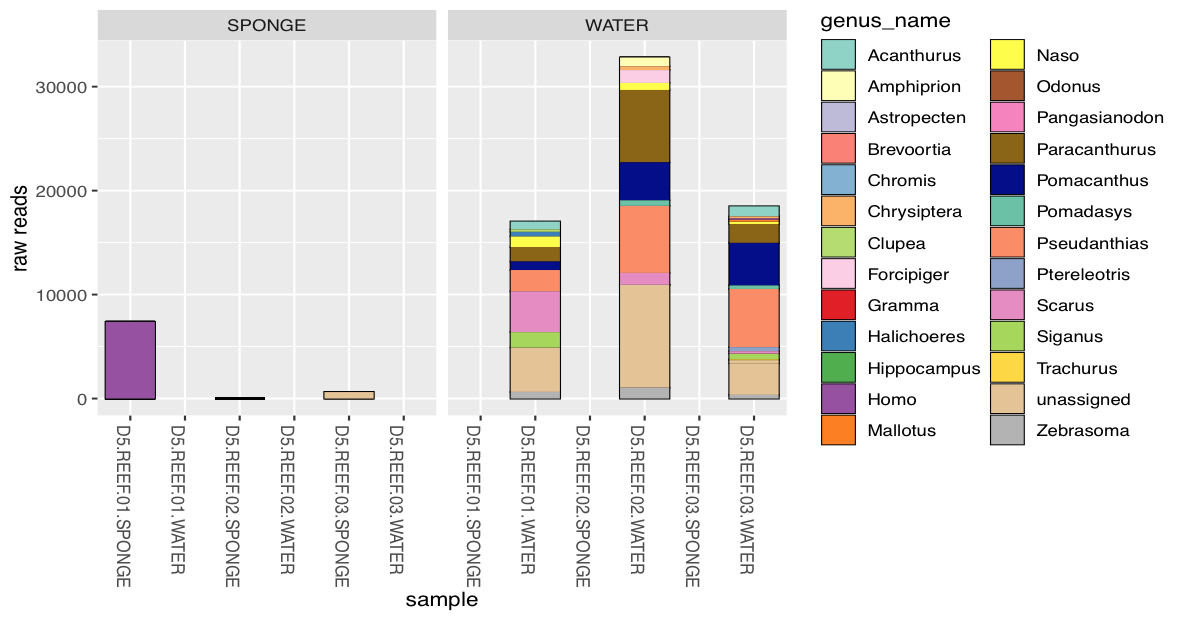
**Figure 4: Bubble plot comparing the efficiency of species detection in each pool.**  The code is the same as figure 3. The sponge may act as an accumulator to continuously collect DNA, so species are sometimes detected later than water samples, and the DNA collected by the pink sponge does not decrease significantly in the short time period.

**Figure 5: Comparison of detection rate among water sample and three sponge species.** Incidence caculate by positive target species in a sample. Read count data is converted to absence/present data. The incidence ranges from 0 to 3. **A.** and **B.** is the comparison of the detection efficientcy of species richness over time; **C.** is the comparison of overall (ignore time) detection efficiency of species richness among sample types.

The efficiency of water sample decreases significantly over time, and followed by cream sponge, while the detection efficiency of pink sponge do not impact by time. There is no difference in the overall detection efficiency among the three of them (water sample, pink sponge, and cream sponge) .

****A B

****C

**Figure 6: Reef experiment.** Sponge samples don’t have positive detections, so I will ignore reef experiment.