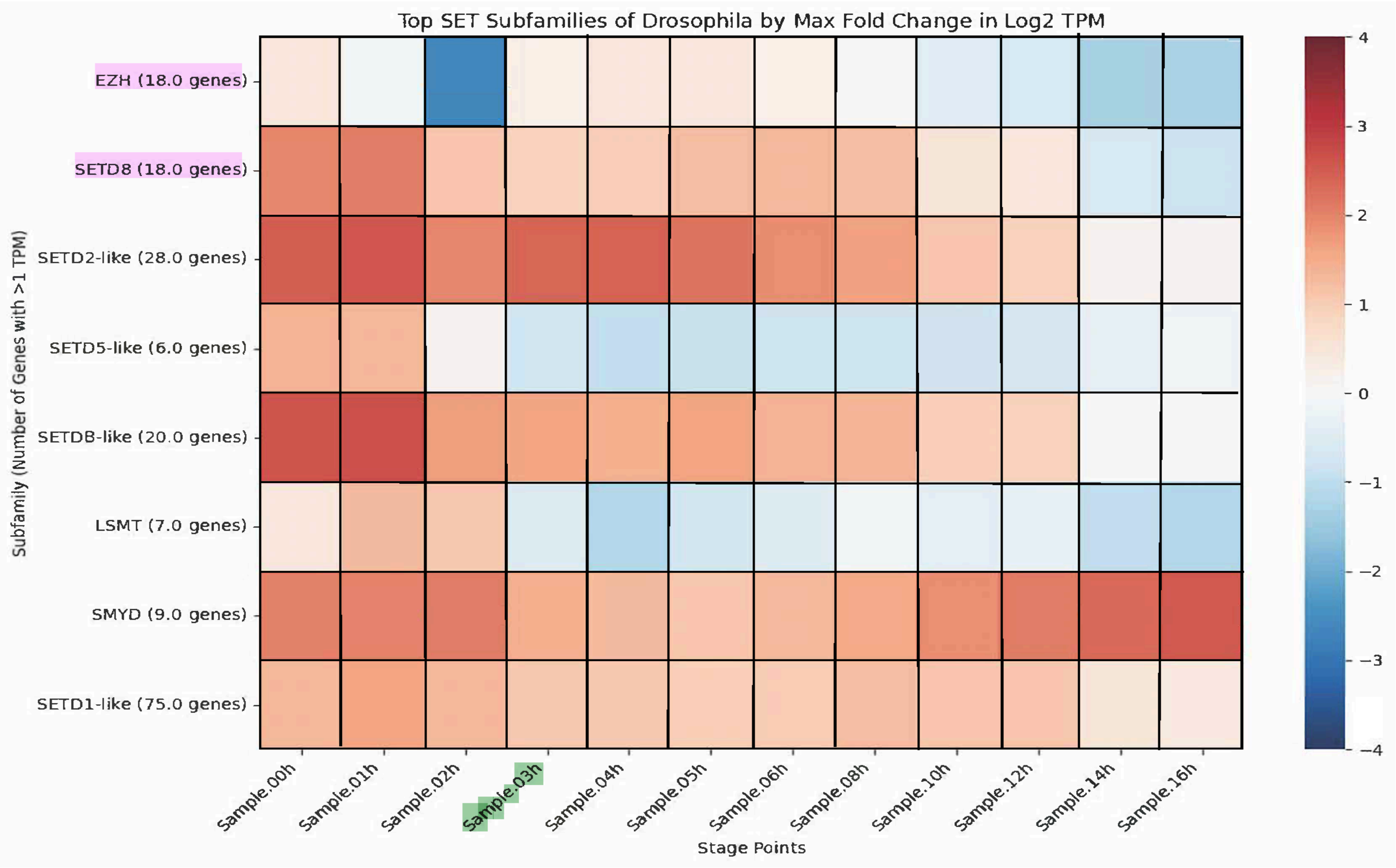

Fig 1. Top SET Subfamilies of Drosophila by Max Fold Change in Log2 TPM. This process combined CSV tables of Drosophila RNAseq data (Comp Bio 1640), SuperFamily CDDIDs (NCBI), Drosophila CDDIDs (BioMart), and SET Subfamilies (NCBI). Once combined, the data was transformed, first by finding unique Subfamily-Gene combinations. For each combination, rows containing TPM values summed together and Log2-transformed (.5 replacing any zero values to prevent calculation errors). For figure generation, Max Fold Change was computed for each row, taking the maximum value of the row and dividing it by the minimum value. These values were centered by subtracting the midpoint of the entire row (calculated by adding together the max and min of the row and dividing by two) to each of the values. The variance was calculated on these mid-point-centered values to highlight the highly variable genes (The top two genes highlighted in pink). The genes were then sorted in descending order of Variance before plotting. The gastrulation period is also highlighted in green to help identify any trends.

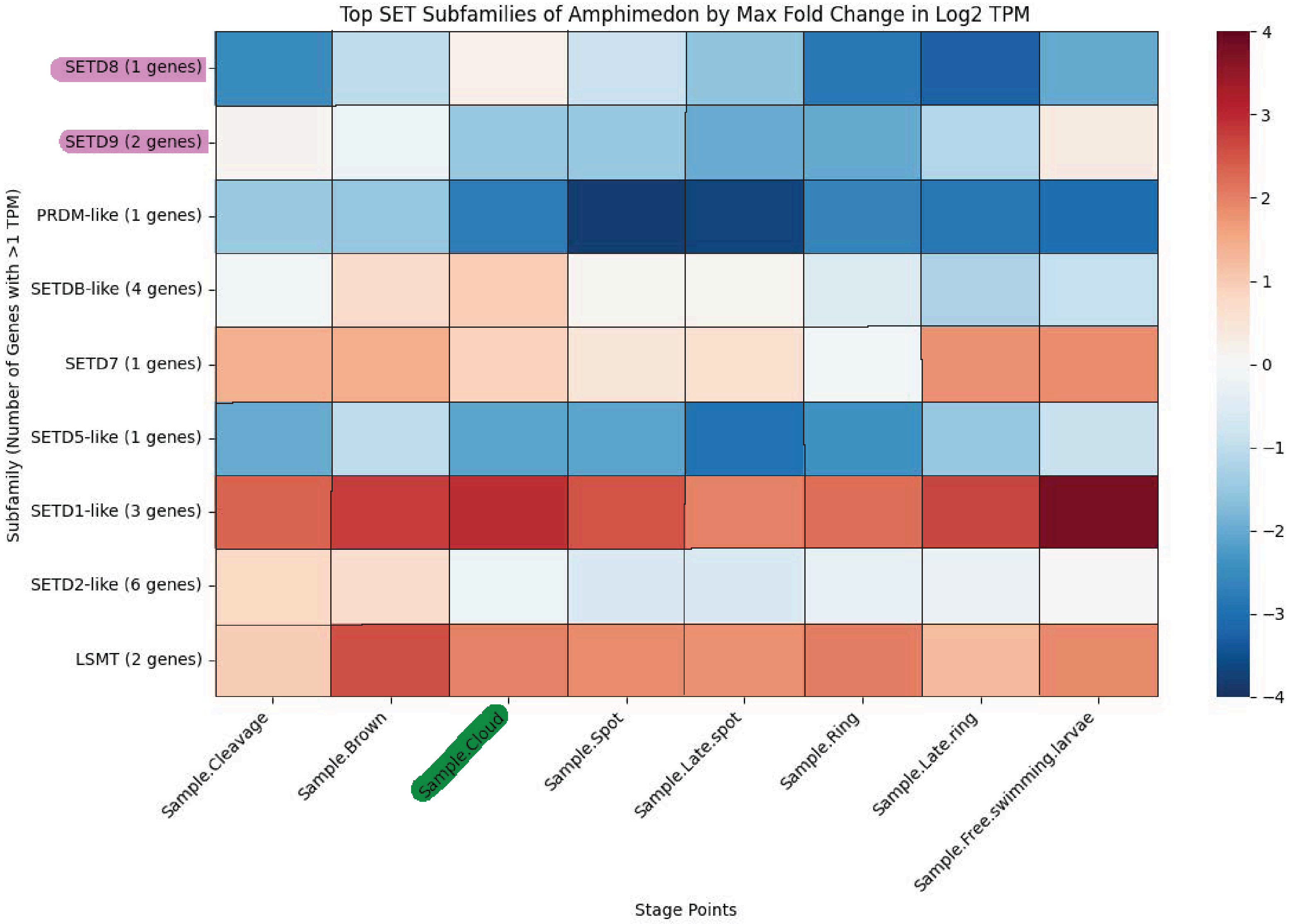

Fig 2. Top SET Subfamilies of Amphimedon by Max Fold Change in Log2 TPM. This process combined CSV tables of Amphimedon RNAseq data (Comp Bio 1640), SuperFamily CDDIDs (NCBI), Drosophila CDDIDs (BioMart), and SET Subfamilies (NCBI). Once combined, the data was transformed, first by finding unique Subfamily-Gene combinations. For each combination, rows containing TPM values summed together and Log2-transformed (.5 replacing any zero values to prevent calculation errors). For figure generation, Max Fold Change was computed for each row, taking the maximum value of the row and dividing it by the minimum value. These values were centered by subtracting the midpoint of the entire row (calculated by adding together the max and min of the row and dividing by two) to each of the values. The variance was calculated on these mid-point-centered values to highlight the highly variable genes (The top two genes highlighted in pink). The genes were then sorted in descending order of Variance before plotting. The gastrulation period is also highlighted in green to help identify any trends.

For this project, I am focusing on the SET Subfamilies and their expression throughout both of the Taxa. I did this by hand since the SET subfamilies had very few actual subfamilies to annotate. I looked up the information on the NCBI website using the superfamily number that was found in the first part of the project. I then created a Google Sheets file that had each Gene's Subfamily and its CDD_ID number. I then transferred this file to a .txt file and merged it with the rest of my tables, using the same code as project one. The difference with this one was that I was now looking at the subfamilies instead of the superfamilies, filtering out ones that did not align with the ones I was observing. However, I used the Top 20 Max Fold Change so i could see how many of these genes are used within each subfamily.