Final Database Relationship Schema

A screenshot of a computer

Description automatically generated

The final database schema includes the relationships between microarray timepoint data collected from papers such as Barreto et al. (2012), Kitagawa et al. (2002), and Thorsen et al. (2007) as well as the data provided from Dahlquist (2018). In standardizing the data for the expression sheets, the header included the author, publication year, yeast strain, log2, and expression such as Thorsen\_2007\_log2\_expression. Within each sheet, there was a master index, ID numbers, the standard names of the genes, and the timepoint data with replicate numbers. Each expression column header was formatted to include the yeast strain, treatment used, concentration and units, log fold change, timepoint, and replicate number such as W303-1A\_arsenite\_1mM\_LogFC\_t15m-1. The degradation rates and production rates data sheets were provided and unchanged, therefore, the column headers included the systematic names, standard names, and degradation or production rates.

The data given from the paper including specific microarray experimental information as well as the paper’s references were included in separate metadata sheets. The sheet headed samples\_to\_data includes columns such as the Sample ID, PubMed ID, control yeast strain, treatment yeast strain, control data, treatment data, concentration of treatment, unit of concentration, time value, units of time, and the number of replicates. The sheet is formatted in such a way that the other expression sheet headers reflect the same format in their headers. The references sheet includes information about the citations of the paper, mentioning the PubMed ID, authors, publication year, title of paper, DOI, and NCBI GEO ID.

With the design of the relationships, each ID of the expression sheets were linked together. In linking the degradation and production rates, the systematic names were linked to the IDs of the expression sheets and to each other. For the samples\_to\_data and reference sheets, the PubMed IDs were linked together and with the final database, the Sample IDs will be linked to the specific timepoint IDs of each of the expression sheets.