README file for calculation of correlations of phenotypes using Biolog's Phenotype Microarray system

Also see: Bourne, CR *et al* (2012) "Classifying compound mechanism of action for linking whole cell phenoptyes to molecular targets" *J Mol Recog* 25:216. doi: 10.1002/jmr.2174

- The calculation method is initiated with the output .csv file from the Biolog Omnilog software. I no longer have access to this software so cannot provide instructions on this output, but an example formatted file is provided: **Omnilog_output_example.csv**

- Note that the calculation method used the **"Avg Height"** parameter after 24 hour incubation at 33°C for all calculations.

- This parameter was captured for each compound at (at least) 3 concentrations in addition to a "plus zero" control.

- The concentrations used for each compound were determined based on the MIC values, such that they were at **0.1, 0.25 and 0.5 the MIC value**. As detailed in the paper, the MIC values were evaluated using the Biolog methodology (also pasted below).

"*S. aureus* ATCC 29213 was propagated on …. Biolog Universal Growth with 5% sheep's blood (BUG-B) …. inoculating fluids were …. IF-0a/IF-10b …. incubation temperature was …. 33°C for 24 h …. the Biolog method the MIC cutoffs were determined on the basis of colorimetric changes in Redox Dye H; and … the colony forming unit for …. for the Biolog method it was 2 to $4x10^6$."

- A bit of a digression about how excel works: in excel, any formula that is created can be updated, such that if the value in the originally referred to sheet changes, it will then also be changed in the referring sheet.

  This means that I can generate data sets from the Omnilog instrument, copy that data set to a generic name, *ie*. "zero_control.csv", "1st_subMIC.csv", "2nd_subMIC.csv", and "3rd_subMIC.csv". Note that these .csv files contain all the PM plate data (*ie.* for PM11 through PM20)

  I can then use a type of Master calculation sheet, such that when I open this file ("Proc_Stds_3conc_SUMdV.xlsx") the formula links are updated with the new values. I can then save a copy of this updated file with a more informative name ("Proc_Stds_3conc_SUMdVTET.xlsx"), and very importantly, *disable link updates!*

- More on the Master calculation sheet: it is separated into one tab per plate and annotated with what is in each well – this should likely be verified as they update plate content in the future!!

  The raw "avg height" values are populated as per well, with a measurement from each of the raw data sets (zero, sub1, sub2, sub3) – note, these are not for biological replicates, but for the series of data from a single experiment! It is very important to perform the

zero, sub1, sub2, and sub3 experiments *at the same time, from the same inoculum, incubated and read in the same way*.

- The avg height is converted to a %growth value (otherwise variations between calculations prevent successful comparisons). The maximum value for all data sets was used at 220, which was empirically determined. The formula is given in the paper, and was taken from a poster presented by the Biolog group where they introduced this technology. It is copied below for convenience:

  "raw numbers were converted to a percent growth using the following formula (Wiater et al., 2007): % growth = [log(height, data point)-log (height, minimum)] / [log(height, maximum in plate)-log(height, minimum)].

- The Master calculation sheet ("Proc_Stds_3conc_SUMdV.xlsx") then carries out a series of different calculations; these are mostly holdovers from when I was trying to figure out the best method forward. This was a very laborious part, because the Biolog calculation method was not available; therefore, I had to come up with different ways to calculate and compare to what Biolog had previously reported. I make no guarantees on my method – "your mileage may vary" – but it worked well for our application at the time.

  What I derived (that I believe is better than Biolog's method) captures the 3D change in shape of the isobologram by:
  - First, calculate the area of shape created by 1 concentration of Biolog's plated compound at 0, 0.1, 0.25, and 0.5x MIC of the exogenously added compound. This used the "surveyor's area" calculation.
  - Then, the area values are subtracted such that the area of the next highest biolog compound is subtracted from the one before it, yielding a value called d(vol). This means that if the volume is smaller, as you would expect for inhibited growth, you will get a negative value for d(vol). Because there are only four plated compound concentrations you will end with three d(vol) values.
  - The three d(vol) values were then summed directly to yield a d(vol) for the Biolog compound in the plate.
  - The same procedure was completed in the opposite direction of the isobologram, that is, the area was calculated for one added compound concentration across the four Biolog plated compound concentrations. This again yielded a summed value for d(vol).
  - **These two d(vol) values were then summed to give the overall change in volume for the given shape of the isobologram ("SUM d(V)") – this is what was used in the correlation analyses.**

- **As a note** – *upon review I have discovered that the axes in every isobologram in these excel sheets has incorrect labels – this does not change the calculation or the results, but is merely a mis-labeling that I cannot be bothered to correct for every plate and well.* What the axes are is "1, 2, 3, 4" correspond to the wells of plated Biolog compound, while "xMIC" refers to the added known antibiotic of unknown test compound.

- **As a note** – for compounds with unknown mechanisms additional xMIC runs were performed, such that additional 4th_sumMIC and 5th_subMIC raw data files were generated.

- On each plate tab there is a summary on the right side that correlates the volumes and changes for the biolog plated, added compound, and potential antagonism. These assessments were not really used for the analysis, but were part of the process of figuring out what the data mean. Further, at the very end of the tabs is a "Sheet 1" – this has examples of perfect data and what the calculations would look like using this method – this was used as a validation / training for the method development.

- The SUMd(V) values were then normalized per added compound to yield fractional values. These are in the excel file **"Database_SUMdVvalues.xlsx",** with raw SUMd(V) on the "format" tab, and the normalized values used for correlations on the "frac" tab.

- An additional filtering step was applied using data from only the "zero" added plates. The rationale was that these should be identical between every run (they were run for every added compound, so 21+ data sets for comparison). Any Biolog plated compound with a growth (using these calculation methods) variation of 20% was removed from all further analysis (including the points with the added compounds). Any plated Biolog compounds that were not contributing to the "fingerprint", that is, the growth did not change, were also removed at this step. (The rationale behind throwing out these points is that the correlations are based on variation; therefore, points with no variation serve as background noise and can make it harder to capture the relationships.) The remaining points are given in the file **"Database_SUMdVvalues_culled.xlsx"**.

- These data points were fed into the excel add-on program "WinSTAT"; however, any statistical package should suffice. The Pearson correlations were calculated and filtered for p-value significance. This is how Figure 1 (and subsequently Figure 2) were generated.

- **I have never tried to use these database values with datasets collected from different times (ie., years) or from different locations. I am not sure how robustly they will scale together. Please drop me a line if you are successful in using these tools with your data!** **cbourne@ou.edu**