Statistical Method Revision

The Plan

The goal is to find one model to rule them all.

The best models have high accuracy and high precision so that will be the first criteria for narrowing them down

Test 1 – Mock Community data only

* Accuracy can be measured by doing a chi squared tested.
* In order to do perform a chi squared we must first calculate expected values. This can be done using the information given in table one of mock community pdf. Can work through systematically the steps of the process to calculate the number of reads that would be expected if every step worked perfectly and every possible read was read. (Note to Chris: Nailing down the exact equation set needed to calculate the expected values will be something we will need to work together on. I think I will need your familiarity with the pipeline.)
* Next requirement for a chi squared test is observed values. These we get from the sequencing results (after being feed through a pipeline). Will require some preprocessing R code to concatenate any taxa represented as multiple rows in the result matrix into one grand total. End result of this should be a data frame where each column is a pipeline and each row represents a unique taxa and each value in the table is the absolute count for the reads of that taxa detected by that pipeline. (Note to Chris: Am I right here to say absolute?)
* Now that we have both our inputs can use base R to perform chi squared test on every pipeline, save the resulting value in a table for use later. It has been discussed that weighted chi square would be beneficial here. This is because a standard chi square test penalizes over and under estimations equally. In practice a model that overestimates is more desirable than a model that underestimates by the same magnitude. The loss of information associated with false negatives is concerned more detrimental than the inclusion of false positives.
* Next we want to measure precision
* High precision, means low variance, means we need to take a look at the variance associated with each pipeline
* This means we need multiple observations per pipeline
* (Note to Chris: Hopefully I have understood the experimental design right if not we can adjust where the replicates are coming from, so it is accurate. In the end all that matters are that we have some form of replicates). A single sample taken of the mock community was sequences 3 times, therefor we received back 3 sets raw data for the mock community sample. Each raw data set will be feed through every pipeline, resulting in 3 outputs per pipeline that estimate mock community composition.
* R can be used to calculate the variance for each set of replicates. We could then take the square root of these variances to instead of standard deviation of the replicates. Not sure if one is beneficial over the other. Something to look into/ask Karl.
* Now that we have calculated a measure of accuracy (chi square test statistic) and precision (variance or standard deviation) we can make a scatter plot depicting the relation between accuracy and precision for all the pipelines at once. Using R plot accuracy of the Y axis and precision on the x. (I am pretty sure this also is arbitrary which is on x and which is on y). For both measurements the lower the statistics the better the model is performing. Therefor the highest performing models will be located in the lower left of our plot as close to the origin as possible.
* We will use this plot to heroically select the top 10 (ideally, this number is flexible based on what we see) models for further analysis.
* Even though we have identified our best preforming pipelines, we still do not know if they produce results we are satisfied with. Further statistical analysis will be performed, exactly we be heavily influenced by what we have to work with.
* Ie:
  + If we were unable to perform a weight chi square as intended early now would be the time to manual determine if any pipelines are prone to underestimation and adjust the ranking accordingly.
  + Create a main effect ANOVA’s to isolate where error is being introduced to see if we can alter the pipelines at all to reduce it.
  + Further adjust the pipeline ranking based off of practical criteria such as run time, CPU requirements, user ease, open source etc.

By this point we should have identified a single pipeline we classify as the best. However, we need to consider that just because this pipeline was the best performer when handling the mock community data, that alone is not a guarantee it will perform well when used on a different community. Ie. Before we commit to this pipeline we want to check for overfitting.

Test 2 – Inclusion of fish tank sequences

* Unlike the mock community there is no way to know taxa are present in the community and at what concentration. We no longer can calculate the expected values required to perform the chi squared test and therefor no longer calculate he accuracy in a way that is comparable to the results of the mock community.
* However, because replicates do exist for this run, we can still calculate the precision of the pipelines when used on the fish tank community in a way that is comparable to the precision attributed to the pipeline for the mock community.
* We will repeat the same precision calculation previously outlined for the mock and compare the results using a scatter plot. Because we are using absence of variance as an indicator for high precision, perhaps will graph these values as 1/var so the lower the variance/higher precision the great the value on the graph.
* We will plot precision attributed to a pipeline when analysing the mock community on the x axis, against the precision attributed to the fish tank community on the y. We probably do not need to do this for all of our pipelines, just the ones we identified as top 10 in the previous test.
* Ideally a positive relationship between precision with the mock community and precision with the fish tank community will be observed. We are using the assumption that if the precision rankings appear to have been conserved between the two communities the accuracy rankings also will have be conserved. We may adjust our ranking of the pipelines if any of them show a noteworthy change in precision between the two communities