1. How is scaling performed?

The flowchart below describes the steps used in the scaling process.

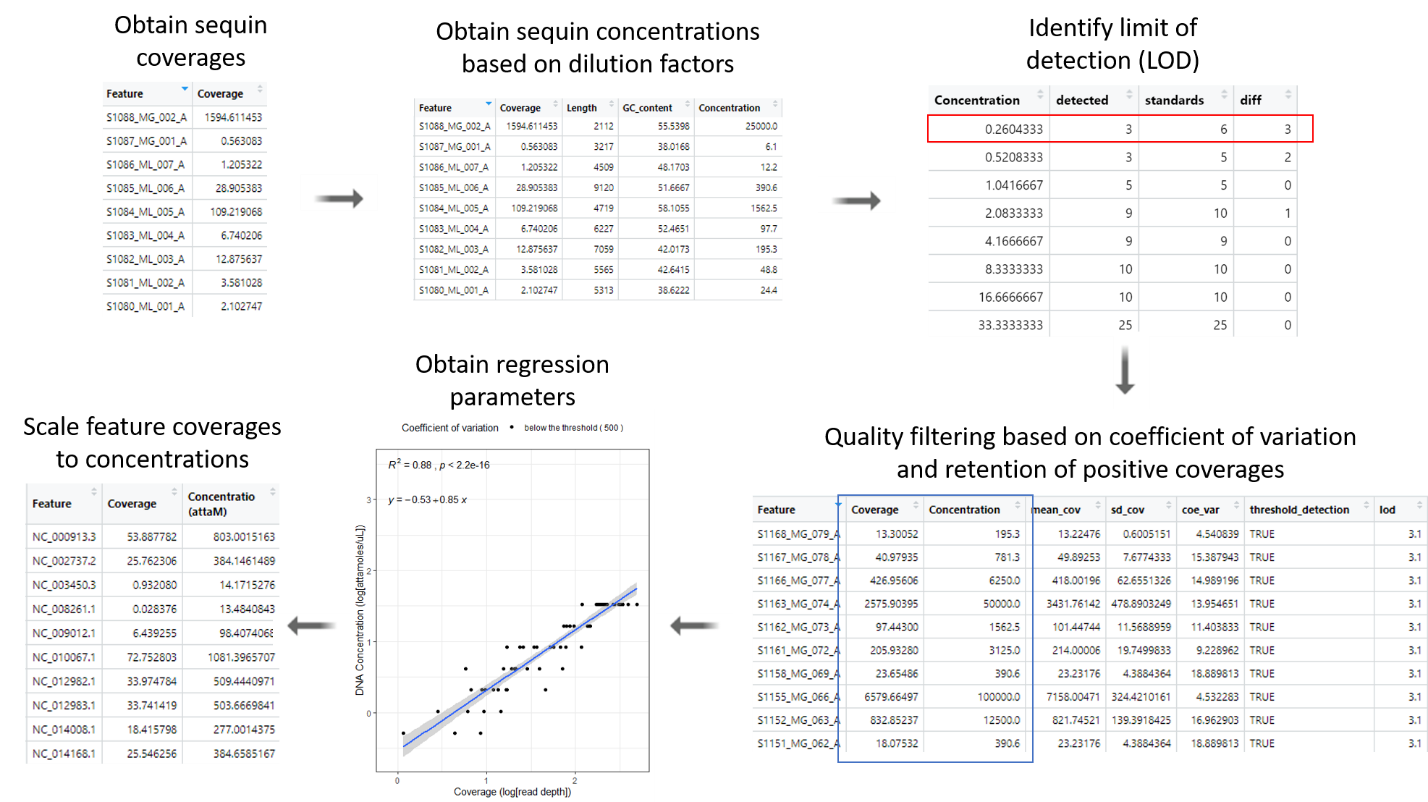


Figure 1: qSIPmg scaling functions in action: Initially, sequin coverages and concentrations are used to identify the limit of detection (LOD) (lowest identified sequin concentration - shown with a red box). Subsequently, quality filtering is performed based on the user-provided coefficient of variation to obtain coverages and concentrations to be used for regression (shown with blue box). Finally, the regression parameters are used to scale coverages of features to obtain absolute concentrations.

1. What is scaling within and beyond the standard curve?

Standard curves are linear regression models with sequin coverages in each fraction as the explanatory variables and sequin concentrations (accounting for dilution rates) as the response variables. As mentioned above, these regression parameters were used to scale the MAG coverages to absolute concentrations.

Some MAGs could have lower coverage compared to the lowest detectable coverage of sequins (coverage corresponding the LOD) in that fraction. If these lower coverages of MAGs were scaled to estimate absolute concentrations, then I called such a scaling approach as “scaling beyond the lower limit of the standard curve”. If these lower coverages were not scaled, they were reported to have zero concentration. I called such a scaling approach as “scaling within the standard curve”. The figure below represents these two scaling approaches. Ideally, we would like to scale within the standard curve and not extrapolate the regression model. But I think it is justified here.

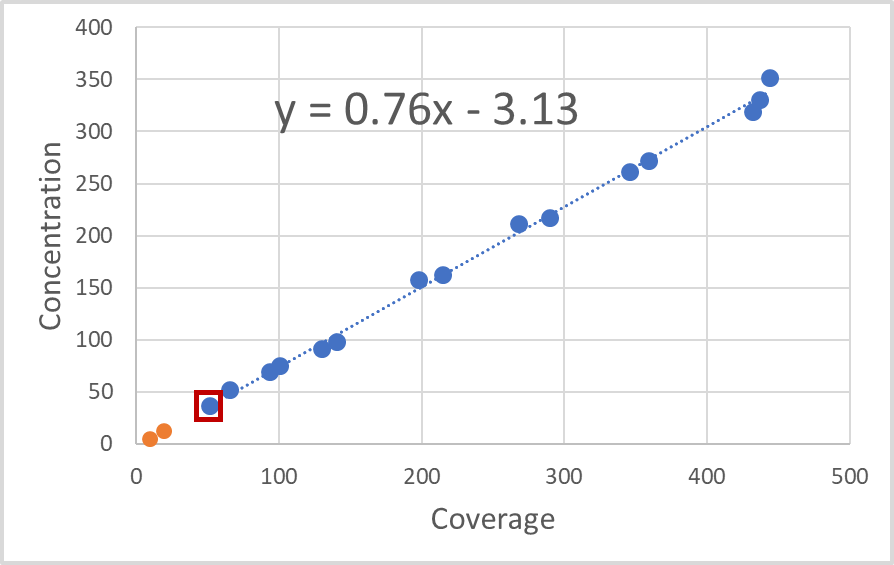


Figure 2: An example of a standard curve from randomly generated data for representation. The blue points represent sequin coverages and concentrations that were used to formulate the regression model. The red square represents the LOD. In the approach “scaling within standard curve”, any MAG with coverage greater than the coverage corresponding to the LOD was scaled to estimate concentration, and any coverage lower than the LOD corresponding coverage (represented as orange points) was reported as zero concentration. In the approach called “scaling beyond the lower limit of the standard curve”, the lower coverages (orange points) were scaled using the same regression model for the rest of the MAGs in this fraction.

1. Justification of the extrapolation

As you know, extrapolation of the regression models is not an ideal path unless justified. But I think it is justified here. Please consider the following graphic for me to illustrate my point.

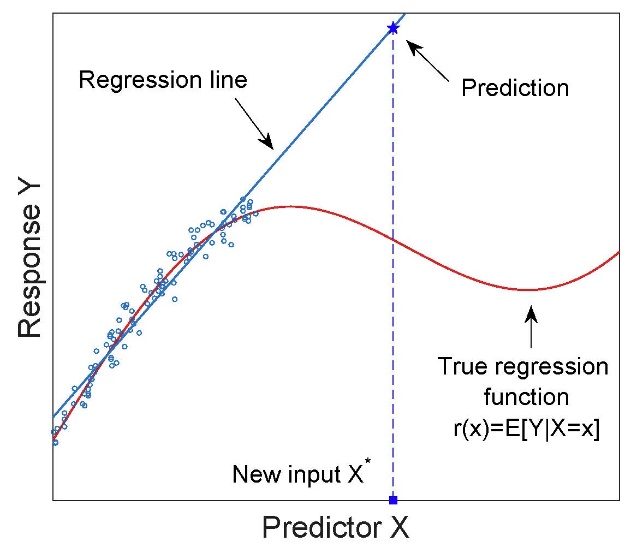


Figure 3

Here, the blue line and points represent a linear regression model that captures the trend throughout the range of determined points. However, the true regression function of this dataset could have been completely different (see the red curve). Thus, if a point beyond the standard curve is used to estimate the response based on the blue regression line, we would have an incorrect prediction because the true regression function of this distribution is not captured by the set of the datapoints used in formulating the blue regression model. So, extrapolation is bad. But it could be justified if the true distribution is captured by the points determined to formulate the regression model. This is true in our qSIP metagenomics case. It is true that coverage and concentration proportionally increase. So, we could use this fact to justify the scaling approach beyond the standard curve.

But even if the distribution is captured by the data used to formulate the regression model, extrapolation could give spurious results. Consider this other graphic that describes how the word “sustainable” becomes unsustainable if extrapolated.

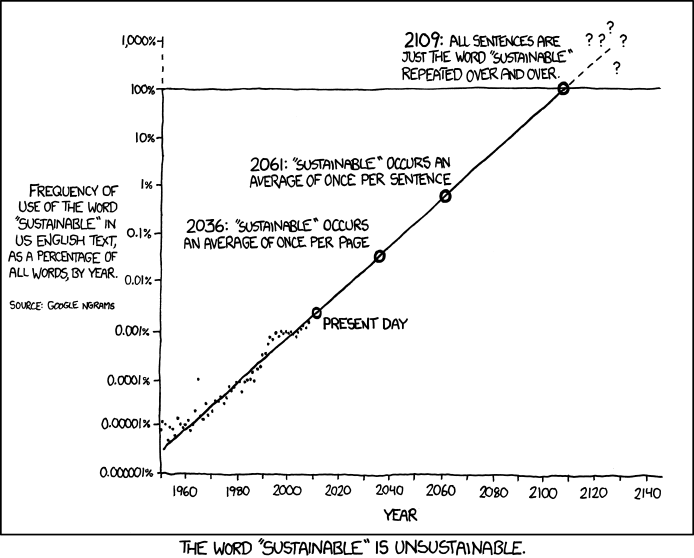


Figure 4

This is a problem if regression models have no meaning beyond the standard curve limits. Consider the orange points in Figure 2. If these MAGs have coverages so low that the regression models predict a zero or negative concentration, then it is a problem. Such an issue can be avoided easily. Currently, there is an insurance in the package scripts for such an event. To avoid breaking the AFE prediction algorithm, the script retains only positive non-zero values. Moreover, the sequin coverages don’t necessarily have to encompass all members, and setting them to zero just because they were of lower coverage means we could lose information about the low abundant members which might be interesting to some users.

But what if there is a MAG with higher coverage than the coverage corresponding to the highest sequin concentration? I still think it is reasonable to estimate concentration because the distribution of the dataset will remain the same (concentration is proportional to coverage) and so long it is not an unreasonable prediction (the concentration being infinite, which would not happen).

Thus, I think it is justified to extrapolate the standard curve in our case.

1. But what about any high false positives in the scaling approach within the standard curve?

One reason why there are more false positives in the approach with scaling within the standard curve is that the estimated weighted mean buoyant density (BD) is altered if low coverages are set to zero. The mean BD is weighted by the abundances, and thus, if there are zero abundances the weighted mean would shift in the direction of the datapoints with a non-zero value. Consider the figure below which shows the abundance across the BD gradient for Bin.227. The profile obtained from scaling beyond the lower limit of the standard curve looks like it is not an incorporator (Figure 5-A). The mean buoyant density for control condition replicates is 1.730 g/mL and for the treatment condition is 1.729 g/mL. Clearly, Bin.227 will not be identified as an incorporator with the qSIP model nor the delta BD method. But if the scaling within the standard curve is done, which reports low coverage values as zero concentration, it confounds the incorporator identification (Figure 5-B). Now the mean BD for control condition is 1.731 g/mL and for the treatment condition it is 1.737 g/mL. Thus, Bin.227 is reported as an incorporator.

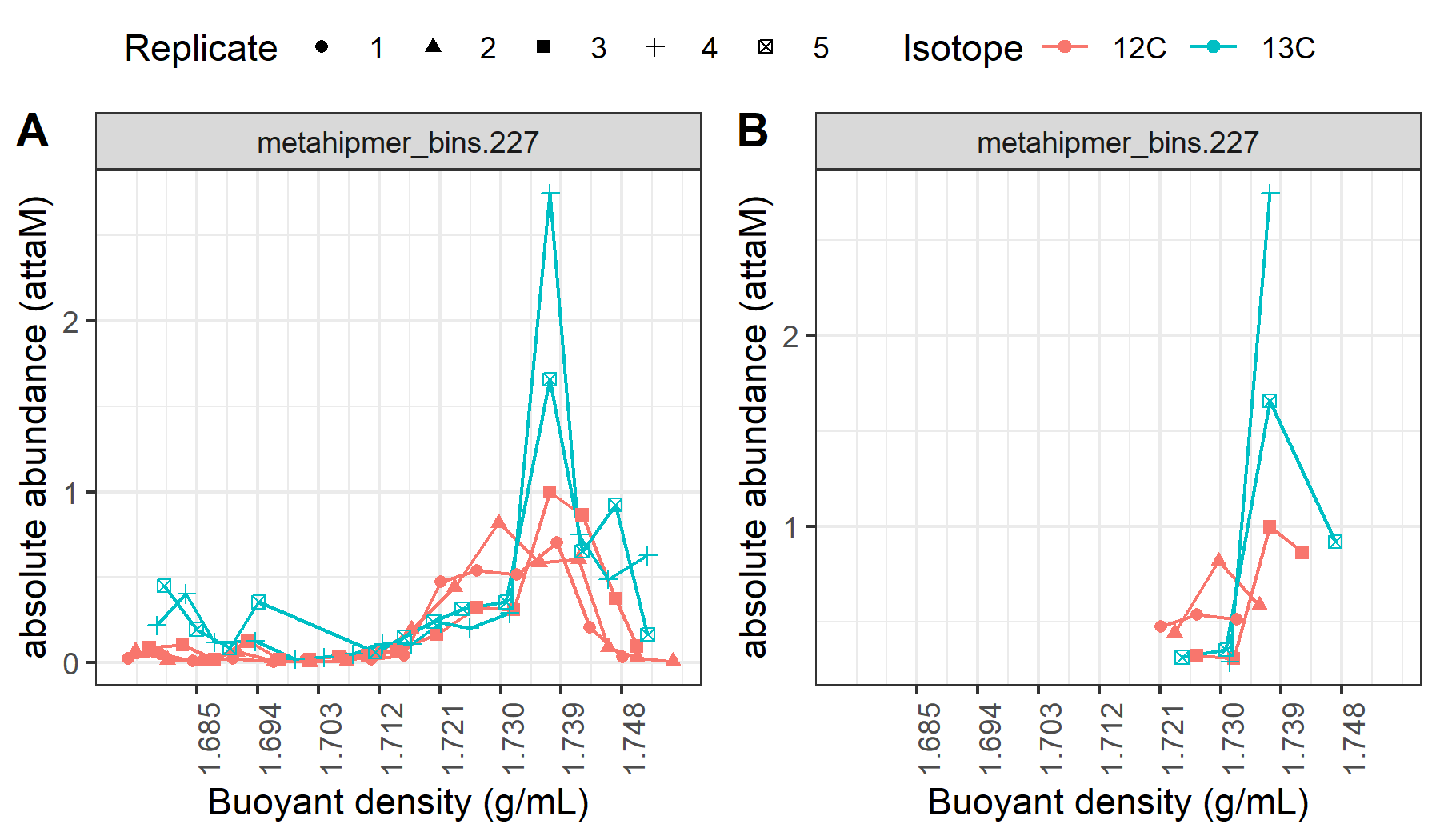


Figure 5: Comparison of abundance profiles between scaling approaches of (A) beyond the lower limit of the standard curve, and (B) within the standard curve.

Such a “pull” of BD towards the higher coverage fractions results in confounding inferences. The above example illustrates how a false positive is reported due to such a scaling approach because of its low coverage. It can be reasoned too that a true incorporator could become a non-incorporator by this method. This scaling approach depends on detectability of sequins, so if in some fractions the LOD associated sequin coverage becomes higher than the low coverage MAGs, then it is not possible to predict which way the mean BD shifts as it depends on the detected abundances in the rest of the fractions. The above illustration is not a one-off example and users need to exercise caution in inferring low coverage of genomic features from scaling approaches.

1. Okay, now what is this LOD-based filter?

To discuss a filtering threshold based on the LOD for each fraction, the following figure is a good motivation.

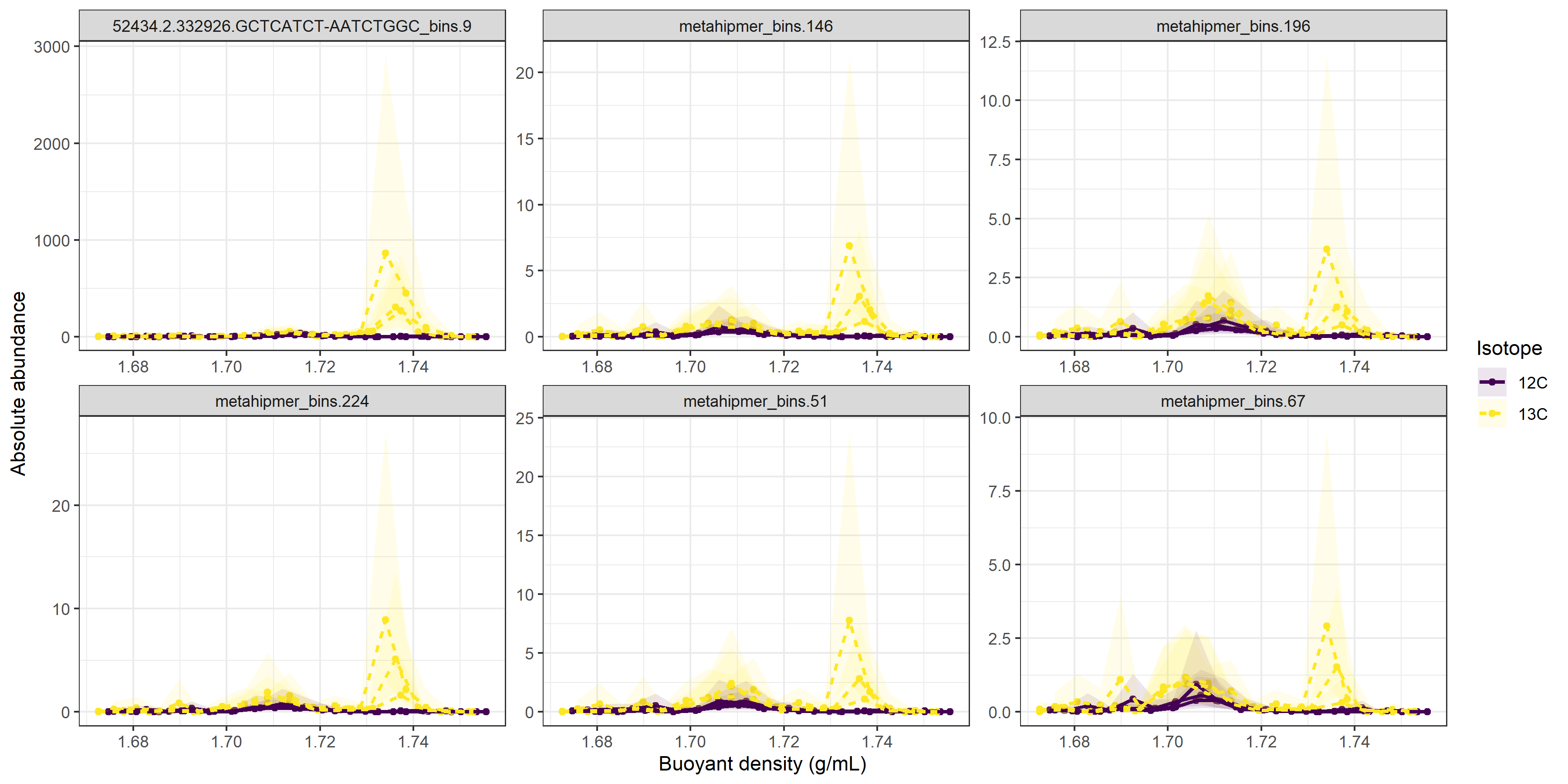
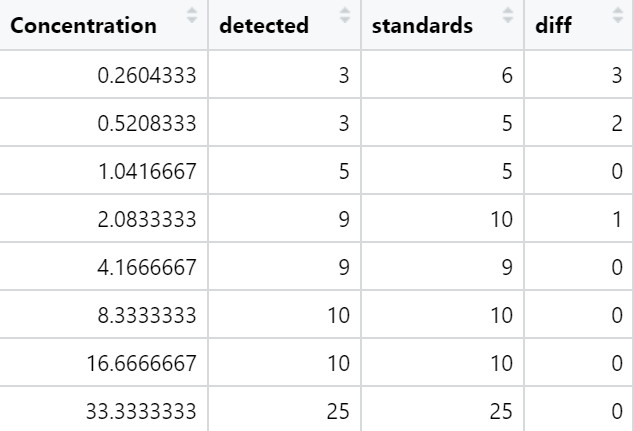


Figure 6: Abundance profile of all identified incorporators in the 40%\_30ng condition.

The ‘long\_name\_bins.9’ is *E*. coli which is isotopically labeled in the 40%\_30ng condition of the published study (Vyshenska and Sampara et al.) and is expected to have 24% AFE. Compared to the true incorporator (*E*. coli) the rest of MAGs have really low abundance (Figure 6). Yet, there is a difference in control and treatment condition peaks, even so at low concentrations, and thus these are reported as incorporators (Figure 6). To avoid these low abundance members, which is basically noise, we could use the LOD-based filter.

The LOD-based filter looks at how many sequins were detected at a concentration level, and compares the count to the number of sequins added. If the LOD% is 50%, then the filter requires that at least 50% of sequins be detected at the lowest concentration group to consider that concentration group in the regression model (Figure 7). The script can iterate to satisfy this LOD-based threshold before sending data to formulate regression models.



LOD at 50% threshold

LOD at 75% threshold

Figure 7: The application of the LOD based filter and identification of different LODs based on set thresholds.

Such a filtering technique did remove the low coverage members, and thus false positives reduced without majorly affecting AFE prediction accuracy for the true incorporator (Figure 8).

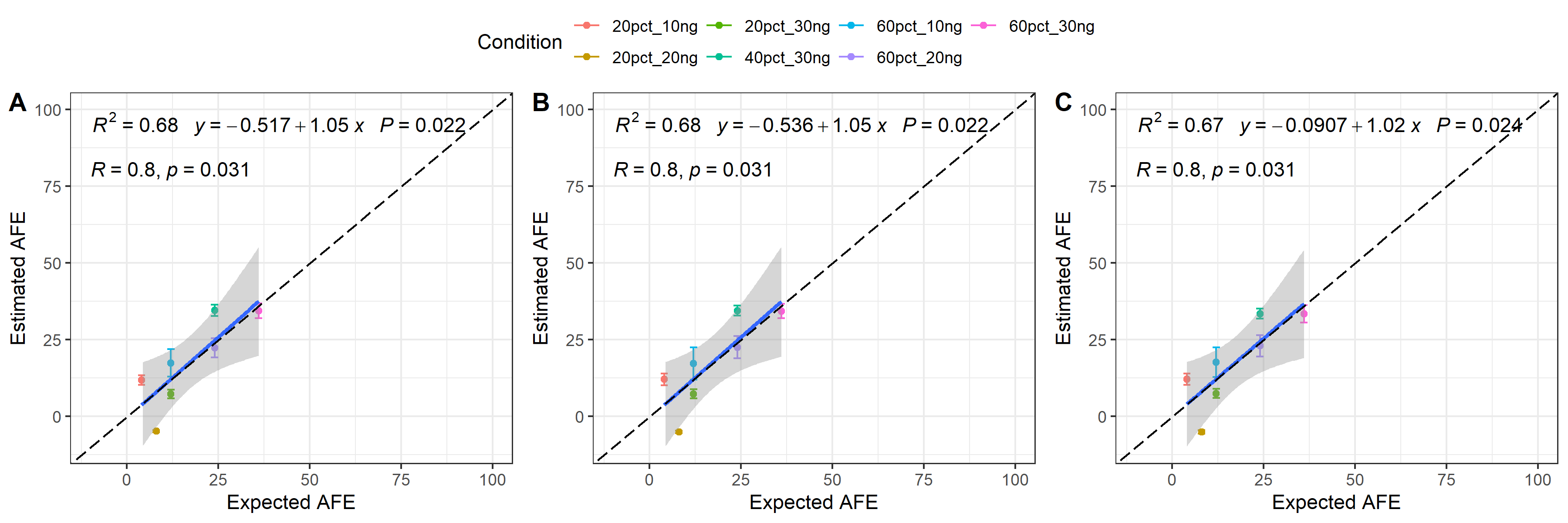


Figure 8: Regression models and correlation between estimated and expected AFE in (A) No filter, (B) LOD50%, and (C) LOD75% cases. LOD filters do not negatively affect AFE prediction.

But the true power of LOD based filtering can be realized only if the low coverage MAGs are set to zero abundance if their coverage is lower than the coverage associated with the LOD. It is reasonable, because that is why the LOD-based filter exists - to filter out noise. This power of LOD-based filtering becomes apparent when we realize that there are no false positives in any condition but for the 20%\_10ng condition with the lowest labeled *E*. coli addition (here there are 21 false positives). With an LOD50% filter, the low coverage noise in the BD profile for the same MAGs in Figure 6 can be mostly removed (Figure 9).

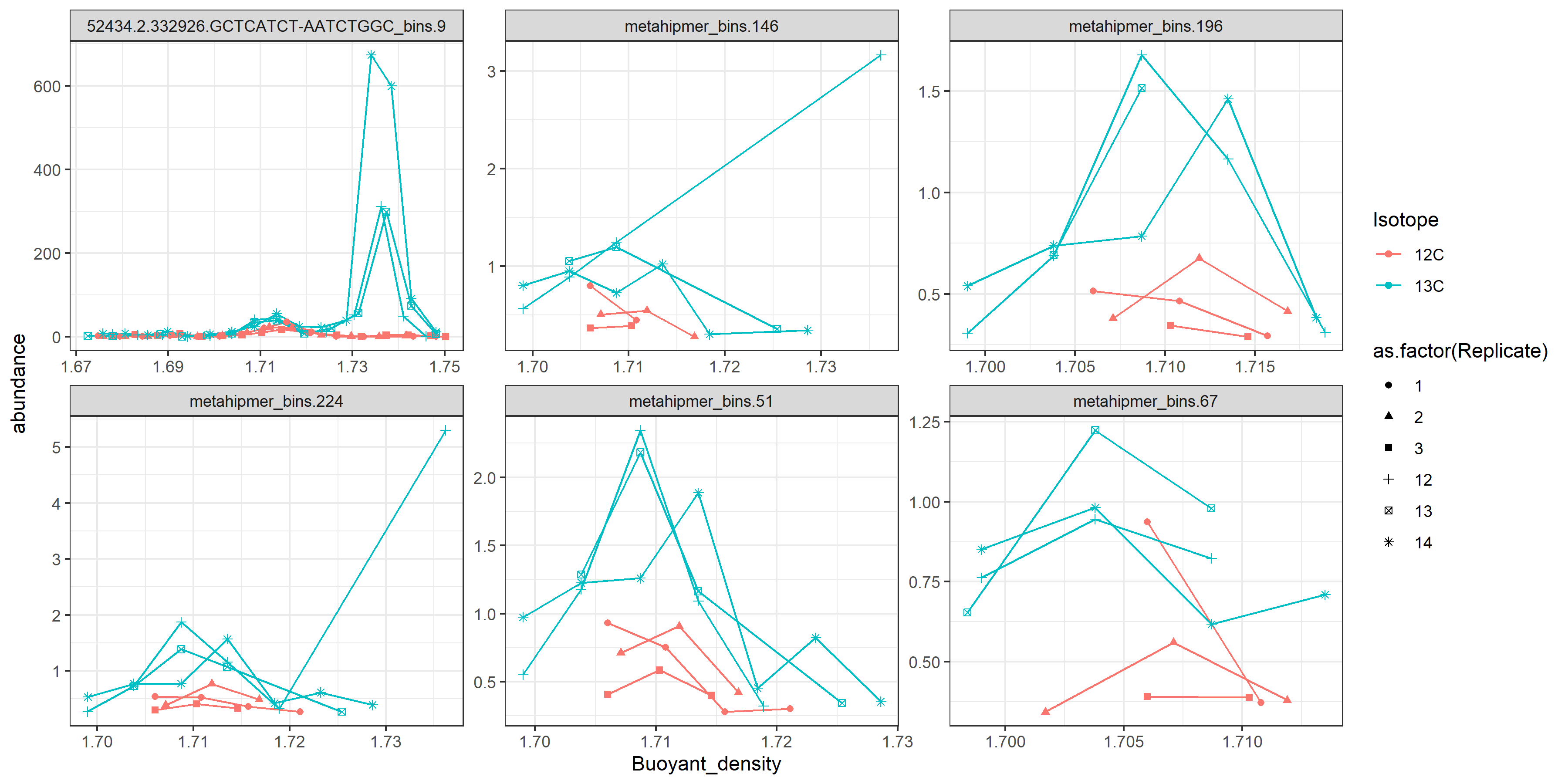


Figure 9: Abundance profile of all identified incorporators in the 40%\_30ng condition with an applied LOD50% threshold for its scaling, and when scaling is performed within the standard curve.

Here, the shift in mean BD for only *E*. coli (“long\_name.bins.9”) is significant to call it an incorporator. I did not realize this before, but this success based on LOD50% filter is probably due to chance. This noise filtering is similar to the no LOD filter applied case, where the scaling was done within the standard curve (Figure 5-B). The same reasoning for possible spurious inferences is valid here too. Here, with the LOD50% filter application, it happened to give a statistically insignificant estimation of AFE due to which fractions got filtered out. It would be impossible to predict how the mean BD would shift, and thus, how AFE would change in a generalized manner.

Thus, although the LOD based filtering makes sense to remove noise and it does not negatively affect AFE of the true incorporator, it alters the number of false positives due to confounding weighted mean BD estimations. Exercise caution if you want to use such a filter.

1. Any other reasons for false positives?

Well, several! The publication discusses in good detail.

1. Can we filter fractions based on a threshold number of fractions to detect in a sample?

What I mean by this approach is this: we decide that there should be X number of fractions detected for a MAG in a replicate across the BD gradient to consider it in the incorporator list.

I don’t think this is a good path forward. A user can design their own method to filter or not, but as a methods suggestion I don’t think it is a good idea. Besides, even if such a filtering threshold is considered, I think that it would be impossible to predict how many fractions are needed to remove noise entirely. Typically, the BD region around the peak would have most of the coverage and as we go in either direction of the peak on the BD gradient, there would be lower coverage. But how can we suggest the number of fractions where one would detect reasonably low coverage? Additionally, if we say that there should be X number of fractions next to the peak in each direction to consider it in the incorporator list, consider the case in Figure 5-B. The BD region with two fractions next to the peak have some abundance in at least one replicate. But because how the BD is “pulled” towards the peak, it results in confounding inferences about AFE estimation. On the other hand, if there are actually low abundance incorporators, it is likely that they would be filtered out due to this threshold as they might have lower coverage than the sequins in certain fractions (a number lower than such a hypothetical threshold for required number of fractions) and we would lose that information.

This makes me think that filtering fractions is not a great idea. What is better is to (a) utilize a threshold for minimum lower confidence interval after bootstrapping to determine isotopic incorporators, and (b) utilize a coverage threshold as discussed above to remove noisy data. See the publication section, “Strategies to improve accuracy of detecting isotopically labeled genomes” for details.

1. How about the DESeq-based method?

Another good way forward (as the paper discusses in detail) is to do DESeq-based filtering first, remove the false positives, then perform AFE estimation with the qSIP model.

In summary:

1. Scaling within the standard curve sets MAG concentrations to zero, which have lower coverage than the coverage associated with the LOD. Scaling beyond the standard curve reports all scaled concentrations regardless of where they fall around the standard curve.
2. Scaling beyond the standard curve can be justified in, and moreover, it is reasonable to apply such a method to avoid confounding results.
3. The true power of LOD-based filtering comes into play only when the scaling is done within the standard curve. It can be performed as an alternative as it highlights the power of using sequins for quality control. Also, a user-specific case might find such a method useful.
4. I think filtering fractions based on a threshold number of detectable fractions for a MAG in a replicate is not a good idea. Instead use thresholds for isotopic incorporator detection and coverage detection limits.
5. DESeq-based methods could be a powerful alternative to reduce false positives.