**MATERIAL S1**Calculation of first-order difference.

The first-order difference gives an approximation of the "elbow" of a graph, the point at which the marginal utility derived from going further along the curve begins to drop off. A common example of this is the variance explained by number of components included graph in principle component analysis. MetaFunPrimer uses a modified approach to finding this point in the graph based on the first-order difference.

The first step in the MetaFunPrimer pipeline is to cluster the input sequences over a range of similarity thresholds. Let and let be the number of clusters found by CD-HIT at similarity threshold . Fig. 2 shows a plot of these values for the *amo*A-AOB genes used in this paper. Then, for each , we calculate the slope of the secant line through and via the formula

Similarly, we calculate the slope of the secant line through and :

We then calculate the first-order difference at by

After calculating for all , MetaFunPrimer uses the value of that maximizes and in the future steps of the pipeline.

While this value is a theoretically optimal value in terms of marginal utility, it is just a suggestion, and the user may have reason to choose a different similarity threshold. In that case, the user can inspect cluster\_counts.tsv, the output of the clustering step, in order to choose their own similarity threshold.