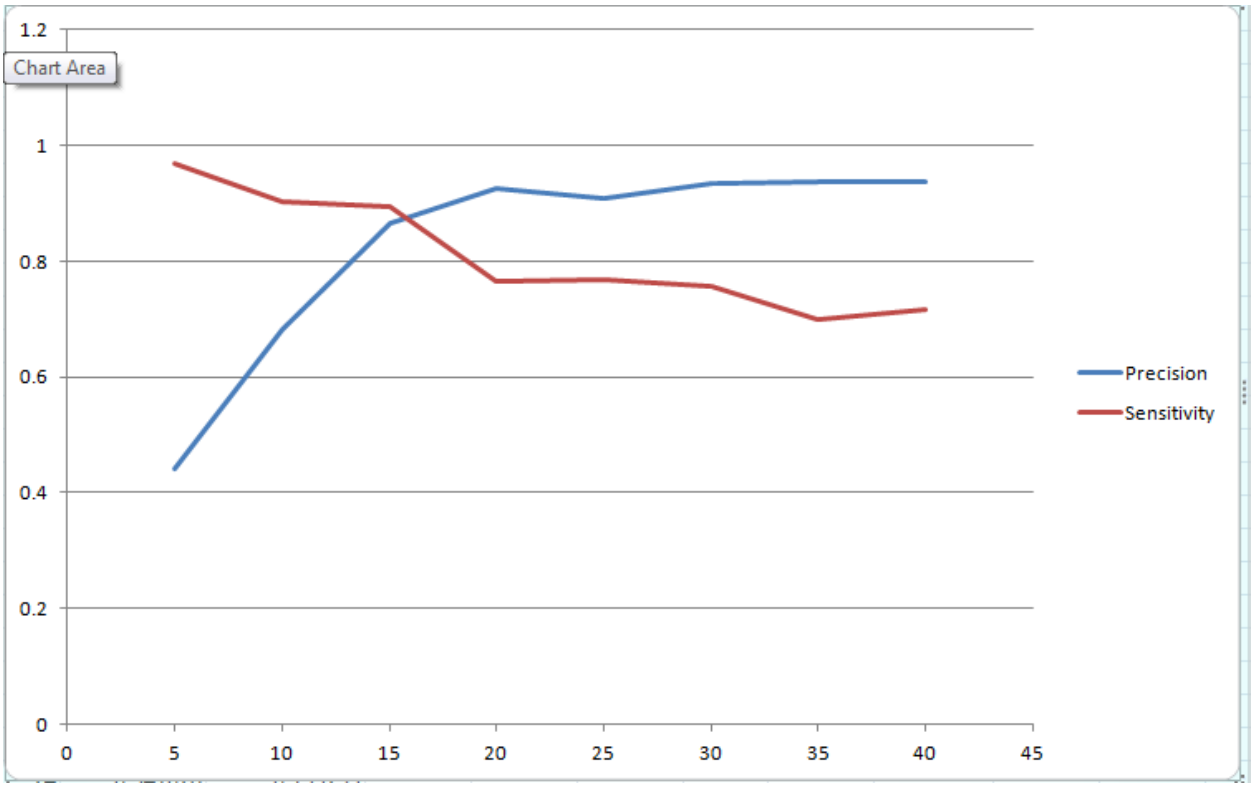
In Metagenomic binning, predicting the correct number of species is very important. In our paper, we use Sensitivity and Precision to measure the binning performance. Assume there are N genomes in the dataset and a binning algorithm outputs M clusters C_i ($1 \leq i \leq M$). Let R_{ij} be the number of reads in C_i which are from genome j and C_j represents genome j_0 when $R_{ij_0} = \max_j R_{ij}$. The overall precision and sensitivity is calculated as:

$$\text{precision} = \frac{\sum_{i=1}^{M} \max_j R_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} R_{ij}}$$

$$\text{sensitivity} = \frac{\sum_{j=1}^{N} \max_i R_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} R_{ij} + \text{number of unclassified reads}}$$

If $M >> N$, the majority of reads in each cluster probably belongs to a single genome and thus precision would be high. However, sensitivity would be low as some genomes are represented by multiple clusters. If $M << N$, some clusters would contain reads from multiple genomes and precision would be low.

The tradeoff between Sensitivity and Precision can be done by fixing different number of species. To show this tradeoff, we modify our MetaCluster5.0 and provide the number of species to the program. We use testing dataset B in our paper and plot the following graph.



The x-axis is the number of species we fixed and y-axis is the value of Sensitivity and Specificity. It's easy to see that precision increases while sensitivity decreases with the number of predicted clusters.