Risks for Specificity and Sensitivity in Metagenomic Classification Experiments

Martin S. Lindner, Bernhard Y. Renard
Research Group Bioinformatics (NG 4)
Robert Koch-Institute, Berlin, Germany

Introduction

Metagenomics is a fastly evolving field in genomics and allows studying the composition of microbial communities far beyond sequencing single genomes. Beyond the enormous possibilities and advances of metagenomics, problems arise if experiments are not correctly designed or results are misinterpreted; e.g. when low concentration contributions of species are missed or incorrect identification occurs.

As Next Generation Sequencing (NGS) experiments remain expensive and repetitions are time consuming, proper experimental design and evaluation and interpretation of experimental results are of high importance. For instance, Stanhope (2010) derived the probability of obtaining at least one contig of certain size in a given setup. Here, we demonstrate how additional computational steps can serve to avoid low sensitivity or low specificity metagenomic classification experiments.

Methods

Sensitivity: Metagenomic experiments allow the parallel analysis of a large number of species. However, one drawback is that most species only account for a very small fraction of all acquired reads. Here, we apply an a priori estimate for the number of required reads for species detection.

In order to detect a species s with relative abundance a in the sample, a minimum number of alignments must be performed in order to be sufficiently sure that s received at least one match. Hence, at least n reads have to be aligned, such that the lower bound of the α-confidence interval of the expected number of matches n_s to s is greater than 0. Assuming that the number of matches is binomially distributed provides an estimate for the number of reads that must be acquired to have confident species detection.

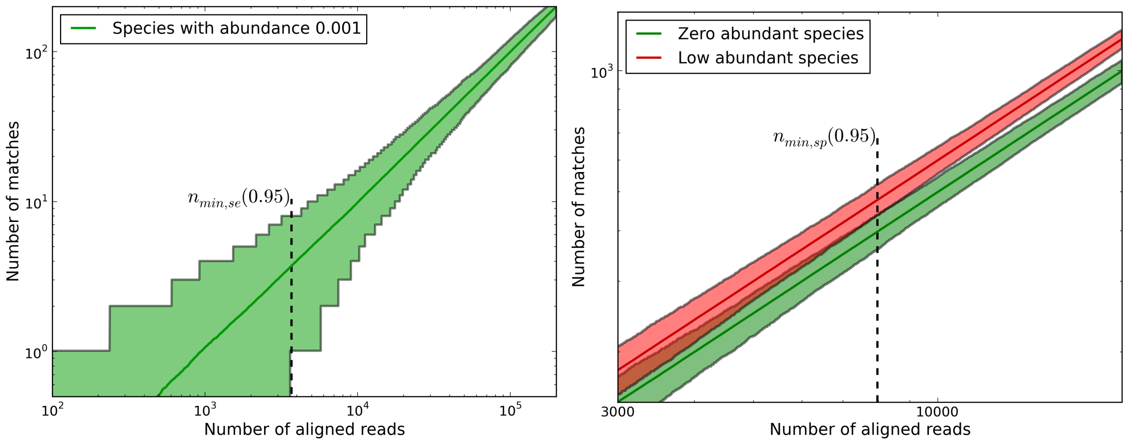
Specificity: Specificity of a metagenomics experiment is at risk when species with high genomic similarity are analyzed, e.g. *Escherichia coli* and *enterohemorrhagic E. coli* (EHEC). Due to the large number of reads which could be aligned to either species, it can be challenging to decide if one or both species are present.

To estimate the probability of an erroneous identification, we take advantage of a normalized distance function between two genomes $d(\cdot, \cdot) \in [0..1]$. Then, we compute a minimum number of alignments n, which is needed to reliably decide whether a species s_2 is present in the sample, even if reads from the similar species s_1 are falsely assigned to s_2. Again assuming a binomial distribution, we derive a lower limit to the number of reads required to reliably distinguish two related species.

Simulation Experiments and Results

Sensitivity: We simulated a large sample containing a species with a relative abundance 0.001 and recorded the number of matches as a function of the number of reads. The simulation was repeated 1000 times and we calculated the mean number of matches as well as the 95%-confidence intervals. Using our derived criterion, we calculated the minimum required number of reads to obtain at least a single match with 95% confidence: $n_{min,se}(0.95) = 3687$. Figure 1a displays the simulated results and shows the high conformance of the predicted number of reads with the simulation.

Specificity: To asses the usefulness of our specificity criterion, we simulated a species s_2 first with zero abundance and then with 0.001 abundance. In both cases, s_2 obtained shared reads from a species s_1 with abundance 0.01 and a genomic distance of 0.5 to s_2. Again, the mean number of matches and the 95%-confidence intervals were recorded. The minimum number of reads required to reliably distinguish the abundant species from the zero abundant species was estimated with our derived formula: $n_{min,sp}(0.95) = 7968$. This theoretical prediction is confirmed by the simulated results, as shown in Figure 1b.



(a) Sensitivity

(b) Specificity

Figure 1: Simulated alignment: (a) shows the number of matches to a reference genome with 95%-confidence interval and the calculated minimum number of reads to obtain at least a single match. (b) shows the number of matches to a zero abundant and a low abundant species in the presence of a high abundant similar species (not shown) together with the calculated minimum number of reads to reliably distinguish the two cases.

Summary

Experimental design can play a vital role for the evaluation and interpretation of metagenomic analyses. For metagenomic classification problems, we derived criteria for the number of aligned reads to ensure the detection of a species (sensitivity) and to distinguish abundant species from zero abundant ones, when the species have highly similar genomes (specificity). We corroborated our considerations by simulated metagenomic experiments, indicating that the derived criteria are highly suitable.

Bibliography

Stephen A. Stanhope. Occupancy modeling, maximum contig size probabilities and designing metagenomics experiments. *PLoS ONE*, 5(7):e11652, 07 2010. doi: 10.1371/journal.pone.0011652.