Detection and Removal of Cross-Contaminations from Transcriptome Sequencing Projects.

S. Nenarokov[1], F. Burki[2], D.J. Richter[3], M. Kolisko[1] and P. J. Keeling[4]

1. Institute of Parasitology, Biology Centre CAS, České Budějovice, Czech Republic
2. Department of Organismal Biology, Uppsala University, Uppsala, Sweden
3. Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Spain
4. Department of Botany, University of British Columbia, Vancouver, Canada

The low cost of next generation sequencing (NGS) allowed affordable and straight forward sequencing of non-model organisms on a large scale. NGS techniques have become a standard for generating transcriptomic and genomic data and it is very common, especially in protistology, for a laboratory to sequence in parallel several different species at a time. Such parallel sequencing commonly leads to a small amount of cross-contamination and even a miniscule contamination is likely to be present in the resulting dataset due to the extremely deep coverage generated by NGS methods. Cross-contamination can arise from both the research laboratory and the sequencing centre. This is especially problematic for transcriptome sequencing projects in which there is no genomic context to confirm the true origin of each assembled sequence. We have developed a software tool for detecting and removing cross-contaminated contigs from assembled transcriptomes. The program uses BLAST to identify suspicious contigs and RPKM values to sort these as either correct or contamination. Through adjustment of the parameters, it also allows for decontamination of species which are taxonomically close or if they are associated by a predator-prey relationship. To demonstrate the effectiveness of our software, we successfully identified cross-contaminations within the ~700 transcriptomes generated by the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) datasets (MOORE foundation) and generated clean datasets.