**Shedding light on the “index hopping” problem in Illumina sequencing technology for amplicon data**

Morais D.K., Baldrian P. and Větrovský T.

Institute of Microbiology of the Czech Academy of Sciences, Videnska 1083, 14220 Prague 4, Czech Republic. daniel.morais@biomed.cas.cz

High-throughput sequencing technologies revolutionized the way biologists deal with data and introduced the biological field to the big-data world. These tools allowed microbial ecologists to assess microbial communities in a level of depth never seen before. However, at the same time, we got analytical and technological problems we have never faced before. For example, the “index hopping” (also called as “tag switching”) events observed in Illumina sequencing technology when using a multiplexed approach. The phenomenon was recognised by the fact that sequence barcodes (also called indexes or tags) were found to be exchanged between samples during the sequencing run forming even such forward and reverse barcode combinations that were not used during amplicon preparation. Furthermore, if a forward or reverse barcode is used for multiple samples during PCR, index hopping can generate incorrect reads with “correct” (existing) barcode combination. Until now, this phenomenon has only been described for 454 and Illumina HiSeq platforms and there are no tools focused on this specific matter. This research aims to supply a tool to identify the switched barcodes and to quantify the frequency of these erroneous events, shedding light on this type of problem. We developed an R script that allows users to identify their sequences by the barcode pair used for each sample, to detect and quantify all the unexpected barcode pairs in a paired-end Illumina sequencing result. The script was completely written using base R functions to avoid unstable dependencies, with most of the functions in vectorised forms to improve speed and memory consumption. The script can be downloaded from <https://github.com/kdanielmorais/Tag_counting_barcodes>. We have inspected five Illumina MiSeq libraries sequenced with the V3 kit that contained 20 barcode pairs and we detected an average of 5% index hopping. For this dataset, the hopping frequency seemed to be affected by barcode sequences or numbers of barcodes used. We believe that a bigger dataset might highlight other important factors and help to explain the molecular mechanisms causing index hopping.