Developments for the statistical analysis of high-throughput sequencing data in infection research

Jochen Kruppa and Klaus Jung
Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover
klaus.jung@tiho-hannover.de

The analysis of high-throughput data has become more and more important in genomic research. Here, we want to present two developments for the analysis of high-throughput data in infection research. Both approaches using high-throughput data generated by next generation sequencing, which has become the state of the art for the analysis of genomic samples. The first one, the gemPlot method concentrates on high-throughput expression data, while the second one, a detection pipeline for viral sequences in reference free host organisms, uses high-throughput sequencing data.

The gemPlot [KJ17] as the 3-dimensional extension of the 1-dimensional boxplot and the 2-dimensional bagplot [RRW99] is a tool for automated outlier detection in molecular high-throughput expression data. Bagplots or gemplots can be applied, separately to the data of each study group, after dimension reduction by principal component analysis. Bagplots and gemplots surround the regular observations with convex hulls and observations outside these hulls are regarded as outliers. The convex hulls are determined separately for the observations of each experimental group while the observations of all groups can be displayed in the same subspace of principal components. The applicability of our method to multigroup data is a clear advantage over other available methods. We provide an implementation of the gemPlot in the R-package 'gemPlot' available from GitHub (`https://github.com/jkruppa/gemPlot`).

The viral detection pipeline is based on raw sequencing reads. The unavailability of a hosts' reference genome is in many cases is a specialty of the infection research in the field of zoonoses. Therefore, reads deriving from the host can not be distinguished from the viral reads by direct mapping against the host reference genome. Here, we present an artificial viral genome approach to circumvent this limitation and ignore the filtering step of the host reads. Further, we use a peptide translated viral reference genome to map the translated DNA reads to achieve a second layer of certainty. Finally, a decoy approach is used to determine the falsely mapped reads and to estimate parameters to judge the correctness of the detection.

References

[KJ17] Jochen Kruppa and Klaus Jung. Automated multigroup outlier identification in molecular high-throughput data using bagplots and gemplots. *BMC Bioinformatics*, 18(1), may 2017.

[RRW99] P J Rousseeuw, I Ruts, and Tukey J W. The Bagplot: A Bivariate Boxplot. *Am Stat*, 53(4), 1999.