Analysis of deep sequencing data to study tumor biology

Wolfgang Raffelsberger[1,*] , Nicodème Paul[1] and Olivier Poch[1] FirstNameA LastNameA[1,2,*] , FirstNameB LastNameB[1,3,4]

1. IGBMC, CNRS UMR7104, Laboratoire de Biologie et Génomique Intégratives, 1 r Laurent Fries, 67404 Illkirch-Strasbourg, France
* Contact author: wolfgang.raffelsberger@igbmc.fr

Keywords: bioinformatics, deep sequencing, SNP statistics

The development of massively parallel sequencing-by-synthesis approaches (such as the Illumina-Solexa and the Roche 454 technologies), also known under the name of deep sequencing, has opened the path for many new applications in biology and medical research. Using such technologies single molecules of DNA (or RNA) can be amplified and sequenced individually at very high throughput. This capacity opens new perspectives in tumor biology since cancer cells acquire during tumor growth novel in a heterogeneous manner mutations, deletions and amplifications in their genome. Furthermore, the deep sequencing approach is very promising, since this provides a uniform platform to compare sequence alterations on the levels of genomic DNA and mRNA.

The mapping of sequences produced from deep sequencing experiments and the statistical analysis of the sequence alterations observed (compared to a reference genome) pose new challenges for users of R. Several packages like Biostrings (Pages et al 2009) and ShortRead (Morgan et al 2009, both on Bioconductor, Gentleman et al 2004) have been developed for running the initial steps of data-analysis, however additional functionalities are needed to study and interpret the characteristics of sequence alterations of inhomogeneous starting material as this is common with cancer biopsies. In this context we are developing a new package dedicated to the reliable identification of sub-populations of sequence alterations form longer sequences (Roche 454 technology). Furthermore, we have developed additional functionalities for the direct comparison of sequence alterations on the levels of genomic DNA and mRNA. This package has allowed us gaining more insight to which degree individual tumors represent actually heterogeneous material on preliminary deep sequencing data.

References

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J.(2004) *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol. 2004;5(10):R80 http://www.bioconductor.org

Morgan M, Lawrence M. and Anders S (2009). *ShortRead: Base classes and methods for high-throughput short-read sequencing data*. R package version 1.0.6

Pages H. (2009), Gentleman R., Aboyoun P. and DebRoy S., *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.10.1