Classification of Coverage Patterns

Stefanie Tauber[1,*], Fritz Sedlazeck[1], Lanay Tierney[2], Karl Kuchler[2], Arndt von Haeseler[1]

1. Center for Integrative Bioinformatics Vienna, Max F Perutz Laboratories, University of Vienna, Medical University of Vienna, University of Veterinary Medicine, Dr.-Bohr-Gasse 9, A-1030 Vienna, Austria
2. Christian Doppler Laboratory Infection Biology, Max F Perutz Laboratories, Medical University of Vienna, Campus Vienna Biocenter, Dr.-Bohr-Gasse 9, A-1030 Vienna, Austria
*Contact author: stefanie.tauber@univie.ac.at

Keywords: Next Generation Sequencing Data, Coverage, Fractals

The advent of DNA sequencing technologies [Metzker (2009)] has brought along an enormous amount of data that still poses a fundamental data-analysis challenge for bioinformaticians and biostatisticians.

When speaking of sequencing data the term 'coverage' is widely used but, at the same time, not well defined. It has to be distinguished between theoretical ('sequencing depth') and observed ('local') coverage. The local coverage can be defined as an integer vector counting per nucleotide the number of reads mapping to the respective nucleotide. In the following the term 'coverage' always refers to the observed local per nucleotide coverage.

In genome resequencing we expect and aim for uniform coverage whereas technologies like RNA-Seq [Ozsolak and Milos (2010)] or ChIP-Seq [Park (2009)] are especially interested in coverage jumps. However, any kind of differential expression analysis relies on a count table containing the number of mapped reads per gene model. This summarization step is not well investigated and its implications on the downstream analysis are not fully understood yet. It is obvious that a summarization value like the sum of reads per gene model is not able to exhaustively capture the underlying coverage information.

Therefore we introduce the fractal dimension (FD) [Kaplan and Glass (1995)] and the Hurst exponent (H) [Peitgen et al. (1992)] in order to distinguish between more or less 'reliable' coverage patterns. FD, as well as H do not make use of any user-defined parameters and are hence free of any ad-hoc heuristics. We propose a re-weighting of the read counts with both FD and H yielding a more reliable count table.

Additionally we show the influence of different mapping strategies on the observed coverage patterns and read counts. This is of course of special interest as any mapping peculiarity propagate to all downstream analysis. We discuss our results on a large Illumina RNA-Seq data set. The host-pathogen interaction of Candida albicans and dendritic mouse cells are investigated by a time course design with three replicates per time point.

We illustrate the entire analysis as well as all up-mentioned methods by means of a *R* package we are developing.

References

Kaplan, D. and L. Glass (1995). *Understanding Nonlinear Dynamics*. New York: Springer.

Metzker, M. L. (2009, December). Sequencing technologies the next generation. *Nature Reviews Genetics 11*(1), 31–46.

Ozsolak, F. and P. M. Milos (2010, December). RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics 12*(February).

Park, P. J. (2009, October). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics 10*(10), 669–80.

Peitgen, H.-O., H. Jürgens, and D. Saupe (1992). *Chaos and Fractals*. New York: Springer.