Classification of Coverage Patterns

Stefanie Tauber,[1] Fritz Sedlazeck,[1] Lanay Tierney,[2] Karl Kuchler,[2] and Arndt Haeseler[1]
[1] CIBIV, Vienna, Austria
[2] MFPL, Vienna, Austria
stefanie.tauber@univie.ac.at

The advent of DNA sequencing technologies 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 or ChIP-Seq 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) in order to distinguish between more or less 'reliable' coverage patterns. The FD does not make use of any user-defined parameters and is hence free of any ad-hoc heuristics. We propose a re-weighting of the read counts with the FD 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 present our results on a 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.