Title & abstract for CAMDA 2011 Conference, Vienna, July 15-16, 2011.

Dealing with the GC-content bias in second-generation DNA sequence data

Terry Speed[1,2] and Yuval Benjamini[2]

[1]Walter and Eliza Hall Institute of Medical Research; [2]Department of Statistics, University of California at Berkeley.

Abstract.
GC-content bias describes the dependence between fragment count (read coverage) and GC content found in high-throughput sequencing assays, particularly the Illumina Genome Analyzer technology. For analyses that focus on measuring fragment abundance within a genome, this bias can dominate the signal of interest. There is no consensus as to the source or shape of the bias; current methods to remove it do not assume a knowledge of the curve shape or scale. In this work we analyze regularities in the GC-bias patterns, and find a compact description for this curve family. It is the GC content of the full DNA fragment, not only the sequenced read, that influences fragment counts. This GC effect is unimodal: both GC rich fragments and AT rich fragments are under-represented in the sequencing results. Moreover, the size of the fragment may interact with the shape and peak of the GC curve. Based on these findings, we propose a new method to calculate expected coverage. This single-bp GC correction and accommodates library, strand, and fragment lengths information, as well as non-uniform bin sizes. We show that it outperforms current approaches in copy-number estimation tasks. These GC-modeling considerations can inform other high-throughput sequencing analyses, such as ChIP-seq and RNA-seq, and illuminate possible causes for the GC-content bias.