Abstract

Determining the methylation state for regions with high copy numbers using second-generation sequencing is challenging because the read length is insufficient, especially when the repetitive regions are long and nearly identical to each other. To resolve these problems, single-molecule real time (SMRT) sequencing shows promise because it is not vulnerable to GC bias, it has long read lengths, and its kinetic information is sensitive to DNA modifications. However, raw kinetic information at a single CpG site contains much noise, and characterizing the DNA methylation for large size genomes demands prohibiting coverage of SMRT reads. Since hypo-/hypermethylated CpG dinucleotides are often contiguous over a long span in vertebrate genomes, we propose a novel algorithm that combines the kinetic information for neighboring CpG sites and increases the confidence in identifying the methylation statuses of those sites when they are correlated. The sensitivity and precision of our algorithm were both of >80% for the genome of an inbred medaka (*Oryzias latipes*) strain within a practical read coverage of <16-fold. With this method, we characterized the landscape of the methylation status of repetitive elements (*e.g.*, the occurrences of ~6-kb-long interspersed nuclear elements (LINEs)) and nearly identical living transposons of 4682 bp long in the medaka genome, which were difficult to observe using bisulfite-treated short reads.