ERNE-BS5: aligning BS-treated sequences by multiple hits on a 5-letters alphabet

Del Fabbro C(1), De Paoli E(3), Prezza N(2), Celii M(3), Morgante M(1,3), Policriti A(1,2)

(1) Istituto di Genomica Applicata, Udine, Italy
(2) University of Udine, Department of Mathematics and Computer Sciences, Udine, Italy
(3) University of Udine, Department of Agricultural and Environmental Sciences, Udine, Italy

Contact: delfabbro@appliedgenomics.org

Motivation

DNA methylation is a widespread epigenetic process that modifies DNA biological properties without changing the underlying nucleotide sequence. DNA methylation consists of the addition of a methyl group to the fifth carbon position of cytosine residues, which produces 5-methylcytosines (5-mC). Methylated cytosines appear primarily in the CG context, although, especially in plants, they may also occur in the CHG and CHH contexts (where H is A, C, or T) at different rates. DNA methylation has important implications for biology and health, being responsible for the regulation of gene expression, imprinting, and silencing of germline-specific genes and repetitive elements. Moreover, methylation differences at intraspecific level or even between homologous chromosomes within the same individual are correlated with patterns of differential gene expression.

The protocol based on bisulfite conversion followed by NGS sequencing (BS-seq) is the gold standard technique used to identify 5-mC-s in a genome at single-base resolution. In the overall process, unmethylated cytosines are converted into thymines leaving methylated cytosines unchanged, thereby providing a device to pinpoint cytosine methylation. This approach poses new computational challenges to read alignment that must be tackled in order to fully and correctly determine the DNA methylation profile of a genome.

In particular, BS-seq read alignment is more complex than standard short-read alignment for a number of reasons:

1) there is a significant increase of the search space, as Watson and Crick strands of bisulfite treated fragments are no longer complementary to each other;
2) there is a reduced complexity of the DNA code since reads belonging to unmethylated sites will be poor in Cs, aggravating the issue of multiple hits;
3) C to T mapping is asymmetric: a T in a BS-treated read can match either a C or a T in the reference, but not vice versa;
4) true evolutionary T to C polymorphisms between samples and reference genomes or between homologous chromosomes cannot be distinguished from C to T substitutions that are caused by bisulfite conversion;
5) more than one BS-seq protocol is available thus affecting the mapping strategy.

Methods

Here we present ERNE-BS5 (Extended Randomized Numerical alignEr - BiSulfite 5), an aligning program developed to efficiently map BS-treated reads against large genomes. ERNE-BS5 is able to align bisulfite reads against a reference dealing with non-symmetric C>T mismatches. Moreover, ERNE-BS5 presents a number of innovative features:

1) we use a 5-letters alphabet for storing methylation information;
2) we use a weighted context-aware Hamming distance to identify a T coming from an unmethylated C context;
3) we use an iterative process to position multiple-hit reads starting from a preliminary map built using single-hit alignments and exploiting between-locus differential methylation as discriminating information. The map is corrected and extended at each cycle using the alignments added in the previous iterations.

Results

We will show that our procedure is able to drastically improve the number of aligned reads at the price of a negligible number of misalignments. Moreover, when the methylation pattern is computed using only single-hit reads, ERNE-BS5 exhibits an accuracy comparable to the other aligners achieving at the same time higher mapping efficiency. In addition, ERNE-BS5 proposes a specific strategy to improve multiple-hit read disambiguation as a preliminary approach to determine methylation patterns in repeated regions. Last, we will illustrate the potential of ERNE-BS5 in using single nucleotide polymorphisms to disentangle DNA methylation signals from two nearly-identical genomes present in a heterozygous individual and reveal differential DNA methylation between homologous alleles.

ERNE-BS5 is based on a new improved version of the rNA aligning software with a more efficient core. ERNE-BS5 is part of the ERNE (Extended Randomized Numerical alignEr) short string alignment package whose goal is to provide an all-inclusive set of tools to handle short reads. ERNE is free software and distributed with an Open Source License (GPL V3) and can be downloaded at: http://erne.sourceforge.net/ .