Methyl-Seq read trimming

Proper trimming of Methyl-Seq reads is critical to accurate downstream analysis. Atropos provides an option to enable automated trimming of Methyl-Seq reads according to the best practices recommended by library construction kit manufacturers or in the literature:

* Reduced-representation bisulfite sequencing (RRBS): RRBS relies on a restriction enzyme (*MspI*) for genome fragmentation. *MspI* leaves a 2 bp overhang that is filled in during end-repair prior to adapter ligation. The filled-in cytosine will not be reflective of the true methylation state, and thus needs to be trimmed away. For reads in which the adapter sequence is detected, Atropos ensures that at least two additional bases are trimmed *after* the adapter sequence is removed.
* Non-directional bisulfite sequencing: Early bisulfite sequencing protocols, including paired-end RRBS and whole-genome bisulfite libraries constructed prior to current-generation protocols (see below), can generate strand-complementary reads whose 5' ends begin with CAA or CGA tri-nucleotides, which are also an artifact of *MspI* digestion. For reads in which the first three 5' bases are CAA or CGA, Atropos ensures that at least 2 bases are trimmed from the 5' end. For non-directional RRBS, the 3' 2 bp of adapter-trimmed reads are removed only if the 5' end does not start with CAA or CGA.
* EpiGnome Methyl-Seq and TruSeq DNA Methylation kits: These kits introduce adapters by tagmentation of bisulfite-converted reads. Trimming of these reads beyond adapter trimming is not required.
* Accel-NGS Methyl-Seq: Accel-NGS (Swift Biosciences) is a recently introduced library construction kit for directional RRBS, WGBS, and other Methyl-Seq protocols. An artifact of adding the adapter sequences is that up to 10 bp of low-complexity sequence are introduced into the 3' end of the template DNA, and thus must be trimmed away. Atropos removes 10 bp from the end of read 1 and the beginning of read 2, as recommended by the manufacturer.

Additionally, in bisulfite mode, Atropos uses an expected nucleotide frequency of 0.33 rather than 0.25 for computing random-match probabilities, since ‘C’ nucleotides are very infrequent. While it would be more technically correct to use nucleotide-specific probabilities for each species and assay type, in practice this level of complexity would have an impact on performance and would be unlikely to change the results substantially, as observed by Sturm et al.