Even though there are a number of excellent short read mapping tools available, e.g. Bowtie (Langmead et al., 2009), these do not perform bisulfite mapping themselves.

2 SOFTWARE DESCRIPTION AND DISCUSSION

Bisulfite libraries are of two distinct types (Chen et al., 2010): in the first scenario the sequencing library is generated in a directional manner, i.e. the actual sequencing reads will correspond to a bisulfite converted version of either the original forward or reverse strand (Lister et al., 2009). In a second scenario, strand specificity is not preserved, which means all four possible bisulfite DNA strands are sequenced at roughly the same frequency (Cokus et al., 2008; Popp et al., 2010).

As the strand identity of a bisulfite read is a priori unknown, our bisulfite mapping tool Bismark aims to find a unique alignment by running four alignment processes simultaneously. First, bisulfite reads are transformed into a C-to-T and G-to-A version (equivalent to a C-to-T conversion on the reverse strand). Then, each of them is aligned to equivalently pre-converted forms of the reference genome using four parallel instances of the short read aligner Bowtie (Fig. 1A). This read mapping enables Bismark to uniquely determine the strand origin of a bisulfite read. Consequently, Bismark can handle BS-Seq data from both directional and non-directional libraries. Since residual cytosines in the sequencing read are converted in silico into a fully bisulfite-converted form before the alignment takes place, mapping performed in this manner handles partial methylation accurately and in an unbiased manner.

A similar approach was demonstrated to work well for single-end reads with the tool BS Seeker, which was developed independently of Bismark (Chen et al., 2010). BS Seeker outperformed earlier generation BS-Seq mapping programs such as BSMAP, RMAP-bs or MAQ in terms of mapping efficiency, accuracy and required CPU time. Even though the principle of both tools is similar, Bismark offers a number of advantages over BS Seeker which are summarized in Table 1. For a test dataset [15 million reads taken from SRR020138 (Lister et al., 2009), trimmed to 50bp, mapped to the human genome build NCBI36, one mismatch allowed], a direct comparison of the two tools returned a very similar number of alignments in a similar time scale [aligned reads/mapping efficiency/CPU time: 9 633 448/64.2%/42 min (Bismark); 9 664 184/64.4%/29 min (BS Seeker)]. Due to the way Bismark determines uniquely best alignments, it is less likely to report non-unique alignments; however, this comes at the cost of a slightly increased run time (for details see Supplementary Material).

Many previous BS-Seq programs were solely mapping applications, which meant that extracting the underlying methylation
data required a lot of post-processing and computational knowledge. Bismark aims to generate a bisulfite mapping output that can be readily explored by bench scientists. Thus, in addition to the alignment process Bismark determines the methylation state of each cytosine position in the read (Fig. 1B). DNA methylation in mammals is thought to occur predominantly at CpG dinucleotides; however, a certain amount of non-CpG methylation has been shown in embryonic stem cells (Lister et al., 2009). In plants, methylation is quite common in both the symmetric CpG or CHG and asymmetric CHH context (whereby H can be either A, T or C) (Feng et al., 2010; Law and Jacobsen, 2010). To enable methylation analysis in different sequence contexts and/or model organisms, methylation calls in Bismark take the surrounding sequence context into consideration and discriminate between cytosines in CpG, CHG and CHH context.

The primary mapping output of Bismark contains one line per read and shows a number of useful pieces of information such as mapping position, alignment strand, the bisulfite read sequence, its equivalent genomic sequence and a methylation call string (Supplementary Material). This mapping output can be subjected to post-processing (Supplementary Material) or can be used to extract the methylation information at individual cytosine positions. This secondary methylation-state output can be generated using a flexible methylation extractor component that accompanies Bismark. The methylation output discriminates between sequence context (CpG, CHG or CHH) and can be obtained either in a comprehensive (all alignment strands merged) or alignment strand-specific format. The latter can be very useful to study asymmetric methylation (hemimethylated or CHH methylation) in a strand-specific manner. The output of the methylation extractor will create one entry (or line) per cytosine, whereby the strand information is used to encode its methylation state: ‘+’ indicates a methylated and ‘−’ a non-methylated cytosine. This output can be converted into other alignment formats such as SAM/BAM or imported into genome browsers, such as SeqMonk, where it can be visualized and further explored by the researcher without requiring additional computational expertise.

### Table 1. Feature comparison of Bismark and BS Seeker

| Feature                                      | Bismark          | BS Seeker         |
|----------------------------------------------|------------------|-------------------|
| Bowie instances *(directional/non-directional)* | 4                | 2/4               |
| Single-end (SE)/paired-end (PE) support      | Yes/yes          | Yes/NA            |
| Variable read length (SE/PE)                 | Yes/NA           | NA                |
| Adjustable insert size (PE)                  | Yes              | NA                |
| Uses basecall qualities for FastQ mapping    | Yes              | No                |
| Adjustable mapping parameters                | 5                | 2                 |
| Directional/non-directional library support  | Yes/yes          | Yes/yes*          |

*Requires library to be constructed with an initial sequence tag (Cokus et al., 2008); NA: not available.

### 3 CONCLUSIONS

We present Bismark, a software package to map and determine the methylation state of BS-Seq reads. Bismark is easy to use, very flexible and is the first published BS-Seq aligner to seamlessly handle single- and paired-end mapping of both directional and non-directional bisulfite libraries. The output of Bismark is easy to interpret and is intended to be analysed directly by the researcher performing the experiment.

**Funding:** Biotechnology and Biological Sciences Research Council (BBSRC).

**Conflict of Interest:** none declared.

### REFERENCES

Chen,P.Y. et al. (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, 11, 203.

Cokus,S.J. et al. (2008) Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452, 215-219.

Feng,S. et al. (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA*, 107, 8689-8694.

Henn,A.A. et al. (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monosomic epigenetic modifications. *Nat. Biotechnol.*, 28, 1097-1105.

Langmead,B. et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10, R25.

Law,A. and Jacobsen,S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.*, 11, 204-220.

Lister,R. et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462, 315-322.

Popov,C. et al. (2010) Genome-wide mutant of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*, 463, 1101-1105.

Robertson,K.D. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, 6, 597-610.