Increased accuracy and speed in whole genome bisulfite read mapping using a two-letter alphabet

Guilherme de Sena Brandine[1] Andrew D. Smith[1*]

[1] Quantitative and Computational Biology, University of Southern California

[*] To whom correspondence should be addressed

{desenabr, andrewds}@usc.edu

December 21, 2020

Abstract

DNA methylation, characterized by the presence of methyl group at cytosines in a DNA sequence, is an important epigenomic mark with a wide range of functions across diverse organisms. Whole genome bisulfite sequencing (WGBS) has emerged as the gold standard to interrogate cytosine methylation. Accurately mapping WGBS reads to a reference genome allows reconstruction of tissue methylomes at single-base resolution. Algorithms used to map WGBS reads often encode the four-base DNA alphabet with three letters by reducing two bases to a common letter.

We introduce another bisulfite mapping algorithm (abismal), based on the novel idea of encoding a four-letter DNA sequence as two letters, one for purines and one for pyrimidines. We show theoretically that this encoding benefits from higher uniformity and specificity when subsequences are selected from reads for filtration. In our implementation, this leads to a decreased mapping time relative to the three-

letter encoding. We demonstrate, using data from multiple public studies, that the abismal software tool improves mapping accuracy at significantly lower mapping times compared to commonly used mappers, with most notable improvements observed in samples originating from the random priming post-bisulfite adapter tagging protocol.

1 Introduction

DNA methylation is a key part of transcription regulation mechanisms in many eukaryotes. This epigenetic mark is characterized by the addition of a methyl group to the fifth carbon of a cytosine (C), which converts it to a 5-methyl cytosine. With few exceptions, heavily methylated promoters are generally associated with epigenomic repression of the gene, and methylation through distal regulatory regions, such as enhancers, tends to be associated with inactivity of the region. In vertebrates, methylation occurs when DNA methyltransferase enzymes recognize cytosines directly followed by guanines (Gs). These CpG dinucleotides are the least frequent of all 16 nucleotide pairs in the genome sequence of most eukaryotes (often comprising only 1-3% of the sequence), but their frequency is significantly higher than average in regulatory regions (Saxonov *et al.*, 2006). DNA methylation is part of biological processes like stem cell fate decision (Smith and Meissner, 2013), genomic imprinting (Li *et al.*, 1993), X chromosome inactivation (Riggs, 1975), and retrotransposon silencing (Molaro *et al.*, 2014). Bisulfite sequencing is the gold standard to analyze DNA methylation (Frommer *et al.*, 1992). Sodium bisulfite treatment converts unmethylated cytosines to uracils while keeping methylated cytosines unchanged. The converted DNA is then amplified in a polymerase chain reaction, which copies uracils as thymines (Ts). Whole genome bisulfite sequencing (WGBS) couples bisulfite treatment with high-throughput sequencing, allowing DNA methylation estimation at single-base resolution and genome-wide coverage (Lister and Ecker, 2009; Zhang *et al.*, 2006). Typical WGBS datasets contain hundreds of millions of short sequences ranging from 50 to 300 bases,

with read numbers constantly increasing as the cost of sequencing decreases. The low relative number of methylated Cs in most organisms causes the frequency of the four bases in WGBS datasets to be largely different than traditional high-throughput sequencing DNA data. Most WGBS reads are T-rich, where Ts account for nearly half of the sequenced bases. If the bisulfite-converted DNA is reverse-complemented prior to sequencing, the most frequent bases will be adenines (As), and reads will be A-rich. We refer to Ts in T-rich reads and As in A-rich reads as bisulfite bases, signifying that they can originate from two possible bases in the DNA (Cs or Ts). Traditional WGBS requires microgram quantities of DNA and generate T-rich reads in single-end sequencing and A-rich reads in the second end of paired-end sequencing. Recent adaptations to the traditional WGBS protocol, developed in attempts to reduce DNA denaturation caused by bisulfite treatment, generate A-rich reads in single-end and in the first end of paired-end sequencing. These post-bisulfite adapter tagging (PBAT) protocols allow WGBS to be performed in nanogram quantities of DNA, often needed in tissues with low cell counts like pre-implantation embryos or metastatic cells (Miura *et al.*, 2012). Some variations of the PBAT protocol use several rounds of random-priming after bisulfite treatment (Miura and Ito, 2015; Clark *et al.*, 2017). These random PBAT (RPBAT) protocols generate randomly sampled A-rich and T-rich reads with the property that corresponding reads in paired-end datasets have complementary bisulfite bases.

Analysis of WGBS data starts by mapping reads to the reference genome sequence of the organism from which the DNA originates. The read mapping problem is defined as finding the substring(s) in the reference with largest similarity to the read, given a specified similarity measurement criterion. Mapping a read to the reference retrieves its most likely location of origin. The conversion of Cs to Ts due to bisulfite treatment makes the computational problem of mapping WGBS reads different than the traditional DNA mapping problem. Specifically, a bisulfite-converted sequence may originate from multiple different four-letter genome sequences with equal probability, so WGBS mappers must allow either Ts or As to match

two possible bases in the genome in order to quantify the similarity between a read and a genome sequence. Due to the large sizes of genomes and the increasingly large number of reads generated by high-throughput sequencers, each read must be mapped with high speed while preserving accuracy. To attain this, mapping algorithms often start with a filtration step, which reduces comparisons to a small set of substrings within the genome that are likely to be optimal candidates. Examples of common filtration steps in read mapping are selecting genome locations that match a fixed read subsequence exactly (Keich *et al.*, 2004), selecting the largest substring in the read where an exact match to the reference exists (Langmead and Salzberg, 2012; Dobin *et al.*, 2013), or selecting the occurrences in the genome of the minimum lexicographic substring of fixed length within the read (Li, 2016; Roberts *et al.*, 2004). In most mapping algorithms, genome locations can be efficiently retrieved through filtration by first indexing the genome, that is, pre-processing the genome sequence to construct data structures that allow efficient filtration. Indexing does not depend on the read to be mapped, so genome sequences are commonly indexed as an initial separate step, with indexing results often being stored on disk and loaded before mapping a set of reads. During mapping, when candidate locations are retrieved in filtration, full comparisons between the read and genome substrings starting at each location are performed, and the best mapping locations are reported if they attain sufficiently high similarity to the read. The time required to map an entire dataset is globally determined by the average time to filter candidates, the number of comparisons performed and the average time required to perform a single comparison, so an effective filtration method is both fast to compute and minimizes the number of comparisons to sub-optimal matches. DNA reads can originate from either of two complementary genome strands, where typically only one strand is available as a reference, and the reverse strand can be inferred by reverse-complementing the given strand. Reads are often mapped to both strands by mapping the original read and its reverse complement to the reference strand, which bypasses the need to have both strands available simultaneously. In the case of WGBS read mapping, the reverse complement of a T-rich read is

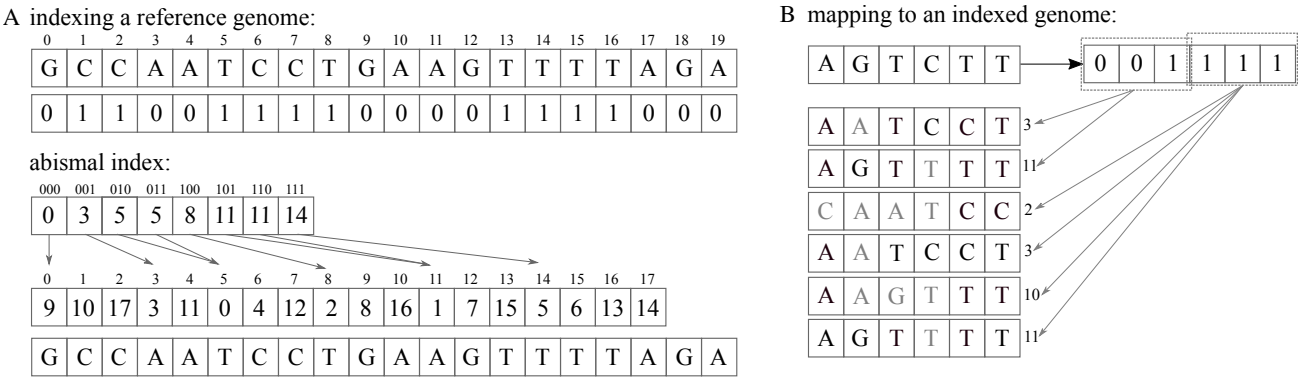

Figure 1: An example of the abismal indexing and mapping algorithms. (A) Indexing of an example genome composed of $M = 20$ bases using $k = 3$ sorting positions. The genome is converted to a two-letter sequence where purines (As and Gs) are encoded as 0s and pyrimidines (Cs and Ts) are encoded as 1s. The index then stores three arrays: a `counts` array of size $2^k = 8$, where each position stores the cumulative sum of k-mer counts lexicographically below the position's binary representation, an `index` array of size $M-k+1 = 18$, where a permutation of the genome positions is stored based on two-letter strings of size k starting at each base, and a copy of the `genome` string, which is kept in the index for subsequent alignments between reads and candidates. (B) Mapping of an example T-rich read with $n = 6$ bases to the reference genome shown in (A). The read is initially encoded in two letters, and uniformly spaced substrings of size $l = 3$ are used to retrieve candidates from the genome using `counts` and `index` arrays. Full comparisons are performed between read and candidates. Mismatches to the input read are represented as gray letters.

A-rich, so mapping a T-rich read to the reverse complement of a strand is equivalent to mapping an A-rich to the original strand.

Several algorithms and software tools have been developed to address the specific challenges pertaining to WGBS read mapping, with diverse strategies that account for bisulfite bases. Bismark (Krueger and Andrews, 2011) converts all Cs, both in the reference and in reads, to Ts, essentially approaching the problem as traditional DNA mapping in a three-letter alphabet, then using the converted reference to map reads using

the bowtie algorithm (Langmead and Salzberg, 2012). An analogous index is created by converting Gs to As, which allows A-rich reads to be mapped. Bsmap (Xi and Li, 2009) leverages the potential specificity benefit of unconverted bases in reads by building a look-up table of all locations of four-letter k-mers of size up to 16. If a given genome k-mer contains Cs, it stores multiple copies of the k-mer location by converting each C as either a C or a T. The resulting table is used in filtration by retrieving candidates based on contiguous substrings from reads and performing Hamming distance comparisons (*i.e.* counting mismatches) between the read and all candidates, but allowing unconverted Cs to only match Cs in the genome. Walt (Chen *et al.*, 2016) adopts a similar framework but uses periodically spaced subsequences within the string (Keich *et al.*, 2004) to filter candidates. Unlike bismark, bsmap and walt limit comparisons to Hamming distance calculations and do not perform full read alignments, although bsmap can be run in a more sensitive mode that allows a single gap of size up to 3. Aligning reads allows mappers to be sensitive to insertions and deletions (indels), but require more computation than calculating Hamming distance (Needleman and Wunsch, 1970; Smith *et al.*, 1981). Indels in reads likely originate from genotype differences between the sequenced DNA and the reference genome. This is often the case in human samples that originate from donors, as the available human reference was sequenced from a small number of individuals and may differ significantly from the donor DNA (Shumate *et al.*, 2020).

Despite their algorithmic and conceptual differences, a common denominator among WGBS mappers is filtering candidates by allowing bisulfite bases to map to two possible genome letters, but requiring all other read bases to match reference letters exactly. If a read subsequence is selected to filter genome candidates based on exact matches, the number of candidates retrieved from the genome will likely increase with the number of bisulfite bases in the selected subsequence, resulting in uneven sensitivity across all possible input reads. Furthermore, mappers that use the three-letter encoding are almost unequivocally implemented by either creating two indices, one for each type of bisulfite base, or by indexing both the read and its reverse

complement, resulting in increased index sizes and often requiring mappers to compromise sensitivity to map reads with feasible time and memory constraints.

Here we introduce a novel WGBS mapping algorithm that leverages the statistical properties of base frequencies in genomes to improve the accuracy of filtration using read subsequences. Strand symmetry in eukaryotic genomes implies that complementary bases are equally frequent in both strands (Kirkpatrick, 2010). A consequence of this symmetry is the property purines (As and Gs) and pyrimidines (Cs and Ts) are also equally represented both before and after bisulfite conversion. This motivated the usage of filtration using a two-letter alphabet encoding, which simultaneously converts purines to one letter and pyrimidines to another letter. Encoding both reads and reference in this alphabet allows filtration in both T-rich and A-rich reads, and mapping sensitivity is independent of the selected sequence under assumptions that are met by real eukaryotic genomes. We show that, compared to the traditional three-letter encoding, mapping reads using the two-letter alphabet increases specificity when a fixed number of bits is available to encode an arbitrary subsequence sampled from a read. We describe theoretical results for fingerprint-based filtration (Karp and Rabin, 1987) in genomes formed by independent and identically distributed (*i.i.d*) letters and show that several genomes have comparable properties to those derived theoretically assuming *i.i.d* sequences. We introduce a mapping algorithm and software tool, named abismal, that presents a new perspective on WGBS read mapping through the two-letter alphabet encoding. Using publicly available datasets, we show that abismal is more accurate than commonly used WGBS mappers and requires less time and memory to map reads.

Dataset	Accession	Species	Phenotype	Protocol	Reads / Pairs	Read length
Ciernia (Vogel Ciernia *et al.*, 2020)	SRR7837058	*H. sapiens*	Prefrontal cortex	Traditional (SE)	174,184,094	100
Manakov (Manakov *et al.*, 2015)	SRR2096734	*M. musculus*	Spermatocytes	Traditional (SE)	308,662,065	100
Mendoza (de Mendoza *et al.*, 2020)	SRR10606701	*D. rerio*	Forebrain	Traditional (SE)	168,367,763	96
Zhang (Zhang *et al.*, 2017)	SRR5015166	*G. gallus*	Breast muscle	Traditional (SE)	139,994,176	125
Shahryary (Shahryary *et al.*, 2020)	SRR12075121	*A. thaliana*	Leaf	Traditional (SE)	76,256,688	151
Bernstein (Bernstein *et al.*, 2010)	SRR1067551	*H. sapiens*	CD56+ ectoderm	Traditional (PE)	168,557,498	100
Decato (Decato *et al.*, 2017)	SRR3897178	*M. musculus*	Whole placenta	Traditional (PE)	124,995,278	90
Kamstra (Kamstra *et al.*, 2018)	SRR5756263	*D. rerio*	Whole embryo	Traditional (PE)	121,650,843	150
Lee (Lee *et al.*, 2017)	SRR5644125	*G. gallus*	Retina	Traditional (PE)	59,768,935	125
Yong (Yong-Villalobos *et al.*, 2015)	SRR2296821	*A. thaliana*	Pit root	Traditional (PE)	21,543,631	90
Miura 1 (Miura and Ito, 2015)	DRR019425	*H. sapiens*	Promega DNA	RPBAT (PE)	69,919,315	100
Bian (Bian *et al.*, 2018)	SRR7461526	*H. sapiens*	CCL23 cell line	RPBAT (PE)	19,986,722	150
Miura 2 (Miura *et al.*, 2019)*	SRR7757863	*H. sapiens*	IMR-90 cell line	RPBAT (PE)	99,575,506	151
Leng (Leng *et al.*, 2019)	SRR9646129	*H. sapiens*	Single-cell 8-cell PGC	RPBAT (PE)	71,902,273	126
Guo (Guo *et al.*, 2015)	SRR2013804	*H. sapiens*	Single-cell 10-week PGC	RPBAT (PE)	91,262,329	125

Table 1: Datasets used for accuracy, speed and memory comparisons between mappers. The asterisk (*) indicates a dataset that was downsampled to 100 million reads prior to trimming to reduce computation time for all mappers.

2 Methods

2.1 Indexing with a two-letter alphabet

Mapping bisulfite sequencing reads typically involves converting the genome into a three-letter sequence, with C→T or G→A to simulate the bisulfite conversion process. Our approach is motivated by the observation that if the reads and the genome are converted into two letter sequences, with both C→T and G→A simultaneously, a match in the two-letter converted sequences is a necessary condition for there to be a

match between the three letter sequences. Moreover, the same two letter encoding works for both T-rich reads and A-rich reads. If this necessary condition is specific enough, and if it can be tested efficiently, the approach may have advantages. Beyond the ability to use the same index for both A-rich and T-rich reads, we present statistical rationale in the next section for why this strategy is justified.

We encode sequences by converting each k-mer to a numerical representation that uses one bit per position, distinguishing purines from pyrimidines. In particular, with $h_2(\mathrm{A}) = h_2(\mathrm{G}) = 0$ and $h_2(\mathrm{C}) = h_2(\mathrm{T}) = 1$, the two-letter fingerprint for a string w is defined as

$$H_2(w) = \sum_{i=0}^{k-1} h_2(w[i]) \times 2^i. \tag{1}$$

Under this scheme, k-mers over the original DNA alphabet map to k-bit binary numbers, and the pre-image of each fingerprint value contains 2^k different DNA strings. This same strategy can be applied in the context of spaced seeds (Ma *et al.*, 2002), encoding ordered but non-contiguous sets of k letters. The advantages of spaced seeds are most pronounced when they span large portions of the query, which renders them far less sensitive to indels. We restrict our analyses to contiguous k-mers.

The two-letter encoding was implemented as part of "another bisulfite mapping algorithm" (abismal), which maps each read independently using two-letter fingerprints retrieved from a set of sampled k-mers that cover the entire read. For read r, we process the k-mers at a fixed set of offsets within r as follows. Given a k-mer w, the set of genomic positions of k-mers sharing the same fingerprint as w are retrieved. Each of these positions is assessed by scoring the alignment of r with the corresponding genomic sequence. The position having the best alignment score is retained as the mapping location for r, and if two locations share the best score, the read is considered to map ambiguously. Most algorithms for mapping short reads fall within this framework. One of the most significant factors in the time required by this framework is the number of full comparisons that must be done for each read. We outline several important implementation choices below, and provide additional details in supplementary methods.

Our implementation uses a direct address table to associate fingerprints with positions in the genome. We use a table to index k-mers with $k = 26$ and then binary search to locate fingerprint matches, initially up to $k = 128$, then reducing to $k = 32$. The direct address table is constructed during genome indexing through a counting sort algorithm (Seward, 1954; Knuth, 1997). Initially, the frequency of all 2^{26} two-letter k-mers is counted in a linear pass through the genome sequence and stored in a count table, which is then transformed into its cumulative sum. A second pass through the genome populates the direct address table by sorting genome position based on their contiguous two-letter k-mer sequences starting at each position and using the resulting count table to group positions with identical k-mers in adjacent locations (Figure 1A). Reads are mapped by selecting k-mers at equally spaced offsets. The best alignment is identified by locating the best approximate match (not allowing indels; Figure 1B) and then computing a banded Smith-Waterman score. Full comparisons are evaluated sequentially so the current best score can be used as a stopping criterion when evaluating subsequent candidate mappings. When too many candidates exist among all selected k-mers, candidates are not compared and the read is considered unmapped.

2.2 Theoretical analysis of the two-letter encoding

Here we analyze the theoretical efficiency of the two-letter fingerprint strategy, defined in equation (1), as a function of the number of bits in the fingerprint. This will be done by comparison with the three-letter encoding previously used in bisulfite mapping algorithms, which reflects the chemical process of bisulfite conversion. In particular, the three-letter strategy encodes letters as h_3 : (A,C,G,T) $\mapsto (0, 2, 1, 2)$, which simply equates T and C. Using h_3, the fingerprint for $w \in \{\text{A,C,G,T}\}^k$ can be constructed as

$$H_3(w) = \sum_{i=0}^{k-1} h_3(w[i]) \times 3^i.$$

The function H_3 evaluates a radix-3 polynomial, while in practice it has been more common to use a radix-4 variation, defined as

$$R(w) = \sum_{i=0}^{k-1} h_3(w[i]) \times 4^i.$$

The two main distinctions between radix-3 and radix-4 are (1) the size, in bits, for the resulting fingerprint value, and (2) the operations required to evaluate the fingerprint function. Note that R can be evaluated using logical operations, while H_3 requires multiplication and modular arithmetic. Since $H_3(w) = H_3(v)$ if and only if $R(w) = R(v)$, the radix-3 and radix-4 encodings are equivalent with respect to accuracy characteristics as fingerprint functions, but when bounded by a common maximum value, the image of the radix-3 encoding contains more elements.

As outlined in the previous section, the time required by abismal and similar algorithms depends on the total number of fingerprint hits, which, for a particular fingerprinting scheme, is proportional to the expected hit rate for a random k-mer. Generally, uniformly represented fingerprints are preferred, that is, with any possible value having equal probability of being observed when a random k-mer is sampled. In the context of mapping through filtration, this not only guarantees that the average number of candidates considered for all reads is minimized, it also implies that the number of candidates considered is not dependent on the selected k-mer.

We assume a random genome sequence of size $M \to \infty$ with *i.i.d* letters, and will address relaxations of this assumption later. We further assume that $\Pr(\text{A}) = \Pr(\text{T}) = p$ and $\Pr(\text{C}) = \Pr(\text{G}) = q$, reflecting strand symmetry in eukaryotic genomes. Importantly, this symmetry ensures that $p + q = 1/2$. We define the expected hit rate of a fingerprint function as the expected relative number of times $H(w)$ is seen in the genome sequence when w is a k-mer sampled from the genome. This emulates the process of selecting fingerprints from a large number of reads that were themselves sampled from the reference sequence, as is the case with real datasets (with the exception of minor contamination). The expected hit rate can be

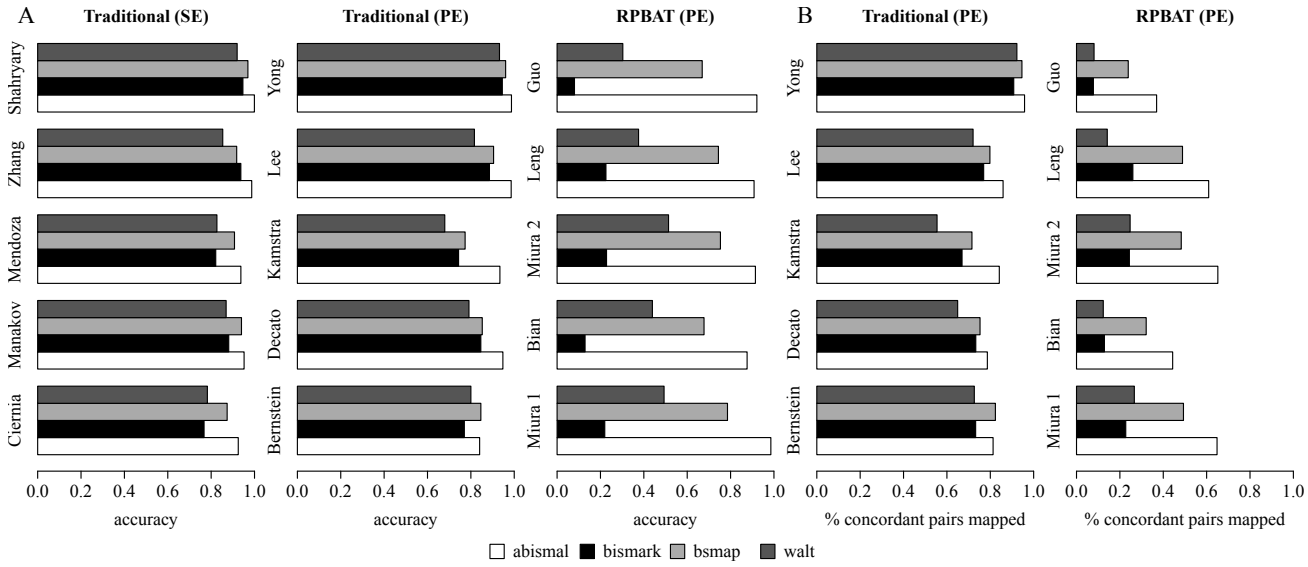

Figure 2: Distribution of the two metrics used to compare mapping results across mappers. (A) Accuracy comparison devised from equation (2) for single-end (SE) datasets (left), paired-end (PE) datasets (middle), both from the traditional WGBS protocol, and RPBAT datasets (right). (B) Number of read pairs with at least one concordant location for traditional PE samples (left) and for for RPBAT samples (right).

calculated as

$$\sum_{w} \Pr(w) \Big(\sum_{w'} \Pr(w') \times I(H(w) = H(w')) \Big), \text{ for indicator function } I.$$

If k-mers w and w' are sampled at random from the genome, this is equivalent to

$$\Pr(H(w) = H(w')).$$

We first consider the two-letter fingerprint strategy H_2 defined in equation (1). Our statistical assumptions on the base frequencies in the genome imply that the expected hit rate is $(1/2)^k$ regardless of the values of p and q, as long as letters are sampled *i.i.d.* Bounding the expected hit rate for H_3 is only slightly more involved.

Proposition 1. *For i.i.d sequences, the expected k-mer hit rate under two-letter fingerprints is $(1/2)^k$.*

Proposition 2. *For i.i.d sequences, the expected k-mer hit rate under three-letter fingerprints is at least* $(3/8)^k$ *and at most* $(1/2)^k$.

Proof. Let w and w' be two independently sampled k-mers, each with *i.i.d* letters. This independence allows the following rearrangement

$$\begin{aligned}
\Pr(H_3(w) = H_3(w')) &= \prod_{i=0}^{k-1} \Pr(h_3(w[i]) = h_3(w'[i])) \\
&= \prod_{i=0}^{k-1} \sum_{j=0}^{2} \Pr(h_3(w[i]) = j)^2 \\
&= \left((p+q)^2 + p^2 + q^2\right)^k \\
&= (1/4 + p^2 + q^2)^k,
\end{aligned}$$

by our assumption that $p + q = 1/2$. The minimal value of $(3/8)^k$ is attained when $p = q = 1/4$, and the maximum value is attained when $p = 1/2$ and $q = 0$. □

If we evaluate fingerprints for k-mers within reads, the specificity is greater for the three-letter strategy, which should be intuitive. However, our capacity to index fingerprint values for efficient retrieval depends on the number of distinct values taken by the fingerprint function. In other words, it is more appropriate to compare the two strategies when the number of fingerprint values is fixed, rather than the number of letters in the fingerprint.

Assume fingerprint values are b-bit non-negative integers. For the two-letter encoding, b letters can be encoded using b bits. For the three-letter encoding, the number of letters that can be represented using b bits is $\lfloor b/\log_2(3) \rfloor$. In this case, the expected hit rate is lower for the two-letter encoding.

Proposition 3. *For fingerprints of b bits, the expected hit rate for the two-letter fingerprint is at most* 0.93^b *times that of the three-letter fingerprint.*

Proof. The ratios from Propositions 1 and 2 give:

$$\frac{\left(\frac{1}{2}\right)^b}{\left(\frac{3}{8}\right)^{\lfloor b/\log_2(3)\rfloor}} \leq \left(\frac{1}{2} \times \left(\frac{3}{8}\right)^{\log_2(3)}\right)^b \leq 0.93^b.$$

□

Proposition 3 indicates that any specificity lost in the two-letter encoding, compared with the three-letter encoding, may be regained by increased uniformity in the distribution of fingerprint values. The property outlined in Proposition 3, however, assumes k-mers are sampled with *i.i.d* letters. Consider a genome containing an extreme abundance of poly-purine sequences. All k-mers sampled from this genome would have the same fingerprint under H_2 but would be distinguishable under H_3. To determine if this happens in real genomes, we calculated the expected hit rates empirically from six reference sequences: *H. sapiens*, *M. musculus*, *D. rerio*, *G. gallus*, *P. troglodytes* and *A. thaliana*, both under the two-letter and three-letter encodings. Our results show, for $2 \leq b \leq 28$, and in all four genomes, the expected hit rate under the two-letter encoding is similar to, but consistently below, the bound derived in Proposition 3 (Supplementary Figure 1), which suggests that the k-mer distribution in the genomes analyzed favor specificity in the two-letter encoding at least as much as in genomes generated by large *i.i.d* sequences. We now examine how an implementation of this encoding affects the accuracy and speed of bisulfite read mapping in real data.

3 Results

3.1 Comparison criteria

The mapped reads are used to estimate a methylation level at each CpG site (or possibly each cytosine) in the genome. The methylation level at a site is the fraction of original molecules (*e.g.* 2 per cell for a diploid sample) carrying the methyl mark at that site. This situation is similar to genotyping a site using mapped

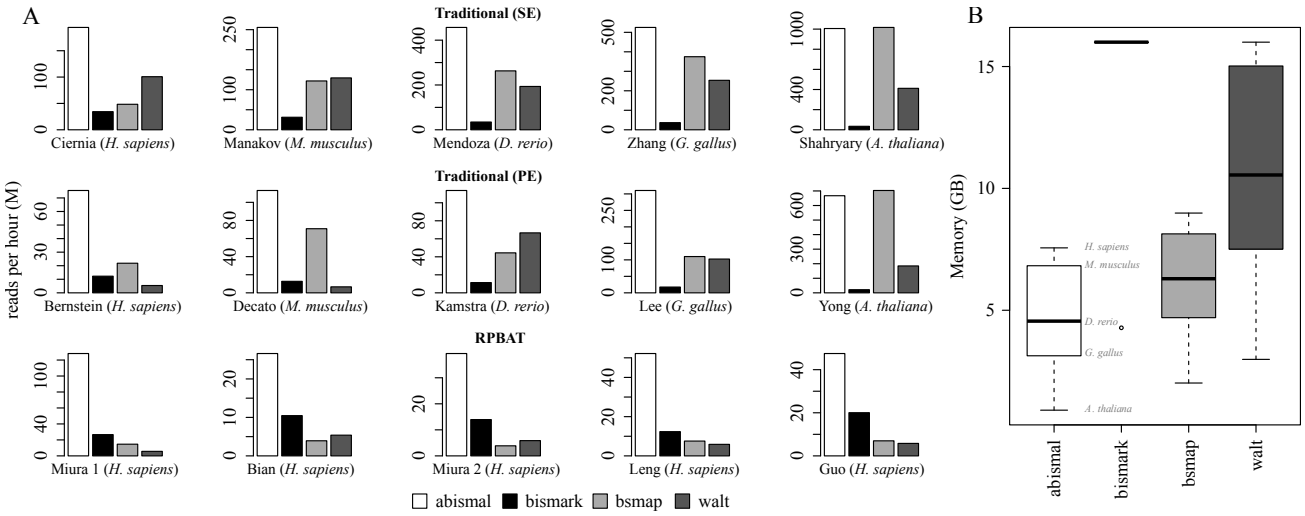

Figure 3: Speed and memory comparison between mappers for traditional single-end (SE), paired-end (PE) and RPBAT samples. (A) Number of reads (in millions) mapped per hour for each dataset selected for comparison. For abismal and bsmap, the total mapping time considered is given by the sum of indexing and mapping times. (B) Maximum memory used to map each paired-end sample shown in Table 1. For bismark, samples that required more than 16 GB for mapping are capped at 16 GB.

whole genome sequencing reads. Unlike genotype, methylation levels may take values beyond the 1.0 and 0.5 frequencies associated with homozygous and heterozygous sites; methylation level may take any value between 0 and 1. Consequently, systematic errors in mapping may lead to incorrect biological conclusions. Below we compare, for multiple mappers appearing in the literature, the frequency with which reads are assigned to their optimal location in the genome.

To perform comparisons, we selected 15 datasets from the public domain. These include single and paired-end datasets from traditional WGBS protocol and from the RPBAT protocol (Table 1). Samples were selected from human and four model organisms (mouse, chicken, zebrafish and arabidopsis), but all RPBAT datasets are from human samples. We show comparison results on two distinct metrics: concordant pairs mapped at least once and estimated accuracy. Concordant pairs are corresponding read ends in a

paired-end datasets that map to opposite strands in nearby locations in the genome (here defined as at most 3000 bases apart). Counting the number of concordant pairs is a comparable metric of sensitivity because abismal, bsmap and walt map read pairs independently, after which these mappers test if any of the best candidates are concordant. Since most species in our tests have large genomes, the smallest with 135 Mbp, it is unlikely that two reads are concordant by chance alone. So counting concordant pairs mapped at least once by each mapper is a sensitivity metric that does not require a ground truth to be available, and mappers would not map concordantly as a consequence of implementation artifacts. We also devised an approximate but more detailed metric that encompasses both sensitivity and specificity of mappers. To define the ground truth status of each read, abismal was run with full sensitivity mode. That is, reads were compared to all candidates retrieved by contiguous fingerprints. This allowed the best possible candidate of each read to be assigned, as well as the edit distance between the read and its candidate location.

Reads are classified relative to a reference in three ways: unmapped, unique or ambiguous, signifying that there are zero, one or many substrings of the genome, respectively, that attain the highest similarity to the read above a minimum desired threshold. Mapping a read uniquely is a sufficient (and often necessary) condition to use it in downstream analyses, so we focus on the accuracy attained by mappers in correctly classifying unique reads. Similarity was measured as the edit distance between the read and the reference, with bisulfite bases considered as matches to both bases they may originate from. All four mappers provide enough information in their output to calculate this value for all reads. Paired-end reads were classified based on the sum of edit distances of both mates.

To measure accuracy, we classify unique mapping reads reported by each mapper with six labels. *Optimal reads* (OR) map to the true correct location from which the read originates. *Sub-optimal* (SR) and *super-optimal reads* (XR) are reads that map to a different location than their true location, but their edit distance is higher and lower than the ground truth, respectively. We define a read to be a *false unique read*

(FR) if it is not super-optimal and if it is classified as ambiguous by the ground truth. False uniques reported by mappers may be considered type 1 errors. *Missed reads* (MR) are reads that are unique in truth but not reported by the mapper. Missed reads originate from type 2 errors. Finally, *dangling reads* (DR) are unmapped by the ground truth but reported as unique by the mapper. Under this classification scheme, every read in a mapper output is assigned to one of five classes: OR, SR, XR, FR or DR. Unique reads not reported are counted as MR. Classes XR and DR correct for cases in which our ground truth may have missed reads or classified them sub-optimally.

Let $\eta(x)$ be the number of reads classified as x. Accuracy μ of each mapper relative to the number of reads in the ground truth is

$$\mu = \frac{\eta(\text{OR}) + \eta(\text{XR}) + \eta(\text{DR})}{\eta(\text{OR}) + \eta(\text{XR}) + \eta(\text{DR}) + \eta(\text{SR}) + \eta(\text{FR}) + \eta(\text{MR})}. \tag{2}$$

The denominator in (2) is the union of all reads classified as unique by either the ground truth or the mapper. The numerator counts both reads whose edit distance is better than the ground truth, or reads whose edit distance is identical and reported as unique by both the mapper and the ground truth. The value of μ is a summary for accuracy that is comparable across mappers, and a more detailed explanation of classification frequencies is provided (Supplementary Table 1).

To compare the time and memory required to map samples, we ran each mapping tool on machines with identical hardware, system software and user environments. We instructed mappers to use 16 threads, which favors mappers that are both fast and make efficient use of parallelism. For abismal and bsmap, the times reported are the sum of the time required to map a read and the time required to index the reference genome the read was mapped to. Unlike bismark and walt, bsmap indexes the genome and maps a sample on a single run. This means that a separate indexing step is not required, but the indexing time is included in the total time required to map a dataset.

3.2 Comparison results

Abismal attains overall higher accuracy (Figure 2A) and total concordant pairs (Figure 2B). We verified that these reads are not a result of a low specificity algorithm by measuring the bisulfite conversion in non-CpG cytosines, as well as the average methylation level in Cs and CpGs for each sample (Supplementary Table 2). The resulting values are similar across mappers, despite differences in reported mapped reads. These observations indicate no large-scale biases related to sequence composition for any of the mappers.

The time comparison shows that abismal is generally faster than bsmap, bismark and walt (Figure 3A). The speed difference seems to increase with genome size, with abismal having significantly higher speed in human samples but nearly identical times to those of bsmap for arabidopsis samples (Supplementary Table 3). We verified that arabidopsis mapping times were similar to the time required to make a copy of the input file, suggesting that run time for this smaller genome is dominated by IO (data not shown). Additionally, the increased sensitivity in RPBAT datasets is attained with much lower mapping times for all samples. The memory requirement for abismal is also significantly smaller than what is required by walt and bismark, and slightly lower than bsmap (Figure 3B). Abismal's indexing times vary from 20 seconds (arabidopsis) to 11 minutes (human; Supplementary Table 3).

4 Discussion

We demonstrated the practical advantage of using a two-letter alphabet encoding to map WGBS reads and its effect on mapping accuracy and speed. Fingerprints under the two-letter alphabet can be stored in one bit per base, allowing larger k-mers to be indexed for a fixed maximum fingerprint size. The uniform representation of both letters in eukaryote genomes makes this encoding a useful method to be incorporated in WGBS read mapping and can be extended to any filtration pattern that selects subsequences from an input read.

The two-letter encoding is not without limitations. There are particular genome distributions in which the expected hit rate for the two-letter encoding increases exponentially with the number of bits relative to the three-letter encoding, even while preserving the assumed property that $p + q = 1/2$. One such example is a genome composed by a large *i.i.d* sequence of purines, followed by an equally large *i.i.d* sequence of pyrimidines. Under two-letter encoding, a fingerprint from a randomly sampled small k-mer would retrieve nearly half of the genome, thus attaining very low specificity, whereas a fingerprint from a three-letter encoding would have low specificity in one half and very high specificity in the other half. This extreme example deviates significantly from the *i.i.d* assumptions, but does not seem to apply to the eukaryotic genomes tested here. On the contrary, our results suggests that there are practical applications of this indexing strategy in many organisms.

Abismal is a free open-source software tool under the GNU Public License (GPL) version 3.0.

References

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., *et al.* (2010). The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, **28**(10), 1045–1048.

Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., Wang, W., Yan, J., Hu, B., Guo, H., *et al.* (2018). Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science*, **362**(6418), 1060–1063.

Chen, H., Smith, A. D., and Chen, T. (2016). WALT: fast and accurate read mapping for bisulfite sequencing. *Bioinformatics*, **32**(22), 3507–3509.

Clark, S. J., Smallwood, S. A., Lee, H. J., Krueger, F., Reik, W., and Kelsey, G. (2017). Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nature Protocols*, **12**(3), 534–547.

de Mendoza, A., Lister, R., and Bogdanovic, O. (2020). Evolution of DNA methylome diversity in eukaryotes. *Journal of Molecular Biology*, **432**(6), 1687–1705.

Decato, B. E., Lopez-Tello, J., Sferruzzi-Perri, A. N., Smith, A. D., and Dean, M. D. (2017). DNA methylation divergence and tissue specialization in the developing mouse placenta. *Molecular Biology and Evolution*, **34**(7), 1702–1712.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1), 15–21.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, **89**(5), 1827–1831.

Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., Yong, J., Hu, Y., Wang, X., Wei, Y., *et al.* (2015). The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell*, **161**(6), 1437–1452.

Kamstra, J. H., Hurem, S., Martin, L. M., Lindeman, L. C., Legler, J., Oughton, D., Salbu, B., Brede, D. A., Lyche, J. L., and Aleström, P. (2018). Ionizing radiation induces transgenerational effects of DNA methylation in zebrafish. *Scientific Reports*, **8**(1), 1–13.

Karp, R. M. and Rabin, M. O. (1987). Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, **31**(2), 249–260.

Keich, U., Li, M., Ma, B., and Tromp, J. (2004). On spaced seeds for similarity search. *Discrete Applied Mathematics*, **138**(3), 253–263.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, **8**(9).

Knuth, D. E. (1997). *The art of computer programming*, volume 3. Pearson Education.

Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**(11), 1571–1572.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4), 357.

Lee, I., Rasoul, B. A., Holub, A. S., Lejeune, A., Enke, R. A., and Timp, W. (2017). Whole genome DNA methylation sequencing of the chicken retina, cornea and brain. *Scientific Data*, **4**, 170148.

Leng, L., Sun, J., Huang, J., Gong, F., Yang, L., Zhang, S., Yuan, X., Fang, F., Xu, X., Luo, Y., *et al.* (2019). Single-cell transcriptome analysis of uniparental embryos reveals parent-of-origin effects on human preimplantation development. *Cell Stem Cell*, **25**(5), 697–712.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, **366**(6453), 362–365.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **32**(14), 2103–2110.

Lister, R. and Ecker, J. R. (2009). Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Research*, **19**(6), 959–966.

Ma, B., Tromp, J., and Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**(3), 440–445.

Manakov, S. A., Pezic, D., Marinov, G. K., Pastor, W. A., Sachidanandam, R., and Aravin, A. A. (2015). MIWI2 and MILI have differential effects on piRNA biogenesis and DNA methylation. *Cell Reports*, **12**(8), 1234–1243.

Miura, F. and Ito, T. (2015). Highly sensitive targeted methylome sequencing by post-bisulfite adaptor tagging. *DNA Research*, **22**(1), 13–18.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, **40**(17), e136–e136.

Miura, F., Shibata, Y., Miura, M., Sangatsuda, Y., Hisano, O., Araki, H., and Ito, T. (2019). Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, **47**(15), e85–e85.

Molaro, A., Falciatori, I., Hodges, E., Aravin, A. A., Marran, K., Rafii, S., McCombie, W. R., Smith, A. D., and Hannon, G. J. (2014). Two waves of de novo methylation during mouse germ cell development. *Genes & Development*, **28**(14), 1544–1549.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**(3), 443–453.

Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, **14**(1), 9–25.

Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M., and Yorke, J. A. (2004). Reducing storage requirements for biological sequence comparison. *Bioinformatics*, **20**(18), 3363–3369.

Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, **103**(5), 1412–1417.

Seward, H. H. (1954). *Information sorting in the application of electronic digital computers to business operations*. Ph.D. thesis, Massachusetts Institute of Technology. Department of Electrical Engineering.

Shahryary, Y., Symeonidi, A., Hazarika, R. R., Denkena, J., Mubeen, T., Hofmeister, B., van Gurp, T., Colomé-Tatché, M., Verhoeven, K., Tuskan, G., *et al.* (2020). AlphaBeta: Computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. *bioRxiv*, page 862243.

Shumate, A., Zimin, A. V., Sherman, R. M., Puiu, D., Wagner, J. M., Olson, N. D., Pertea, M., Salit, M. L., Zook, J. M., and Salzberg, S. L. (2020). Assembly and annotation of an Ashkenazi human reference genome. *Genome Biology*, **21**(1), 1–18.

Smith, T. F., Waterman, M. S., *et al.* (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1), 195–197.

Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, **14**(3), 204–220.

Vogel Ciernia, A., Laufer, B., Hwang, H., Dunaway, K., Mordaunt, C., Coulson, R., Yasui, D., and LaSalle, J. (2020). Epigenomic convergence of neural-immune risk factors in neurodevelopmental disorder cortex. *Cerebral Cortex*, **30**(2), 640–655.

Xi, Y. and Li, W. (2009). Bsmap: whole genome bisulfite sequence mapping program. *BMC bioinformatics*, **10**(1), 1–9.

Yong-Villalobos, L., González-Morales, S. I., Wrobel, K., Gutiérrez-Alanis, D., Cervantes-Peréz, S. A., Hayano-Kanashiro, C., Oropeza-Aburto, A., Cruz-Ramírez, A., Martínez, O., and Herrera-Estrella, L. (2015). Methylome analysis reveals an important role for epigenetic changes in the regulation of the Arabidopsis response to phosphate starvation. *Proceedings of the National Academy of Sciences*, **112**(52), E7293–E7302.

Zhang, M., Yan, F.-B., Li, F., Jiang, K.-R., Li, D.-H., Han, R.-L., Li, Z.-J., Jiang, R.-R., Liu, X.-J., Kang, X.-T., *et al.* (2017). Genome-wide DNA methylation profiles reveal novel candidate genes associated with meat quality at different age stages in hens. *Scientific Reports*, **7**, 45564.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., *et al.* (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell*, **126**(6), 1189–1201.