BSImp: Imputing partially observed methylation patterns

Ya-Ting Chang, Ming-Ren Yen[1] and Pao-Yang Chen[1]
[1]Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan
contact: ytchang.sabrina@gmail.com

Contents

1 Overview

Imputation recovers partially observed methylation patterns for the analysis of methylation heterogeneity at a large proportion of regions genomewide and also estimates methylation levels accurately.

2 Method

Considering methylation patterns formed by methylation statuses of multiple successive cytosines of the same reads, there is usually missing values of methylation statuses within a window of fixed size (number of cytosines). We assume that the pattern of methylation is similar for cells within a population and that the behaviour of cells or the methylation statuses of a cell at a given position can be predicted by those statuses nearby and cells nearby; therefore, using law of total probability, let the methylation status of a cytosine at a position j for read i be m_{ij}, then the probability of m_{ij} being methylated, or 1, is

$$\begin{aligned} p(m_{ij}=1) = & \, p(m_{-i,j}=1|m_{i,-j}=s_1)p(m_{i,-j}=s_1) \\ & + p(m_{-i,j}=1|m_{i,-j}=s_2)p(m_{i,-j}=s_2) \\ & + \cdots + p(m_{-i,j}=1|m_{i,-j}=s_n)p(m_{i,-j}=s_n) \end{aligned}$$

where p_i are subpatterns of complete patterns within the same window, or methylation patterns at positions other than j and $p(m_{-i,j}=1|m_{i,-j}=s_u)$ is the observed probability of cytosines being methylated at position j given subpattern $m_{i,-j}$ within the window is like s_u.

The reads eligible for imputation is specified to be those missing at most 1 methylation status within the window. Since m_{ij} is the only missing value in the window for the same read, $m_{i,-j}$ must equal to one of $p(m_{-i,j} = 1|m_{i,-j} = s_u)p(m_{i,-j} = s_u)$ where $u \subset \{1, 2, \ldots, n\}$. However, if the subpattern is not observed, or there is no complete pattern with subpattern that resembles $m_{i,-j}$, it is taken as the methylation level at position j, or $p(m_{tj} = 1)$ for all reads t that are observed at position j. An illustration of the eligibility of reads for imputation and a possible imputation result can be found in Figure 1.

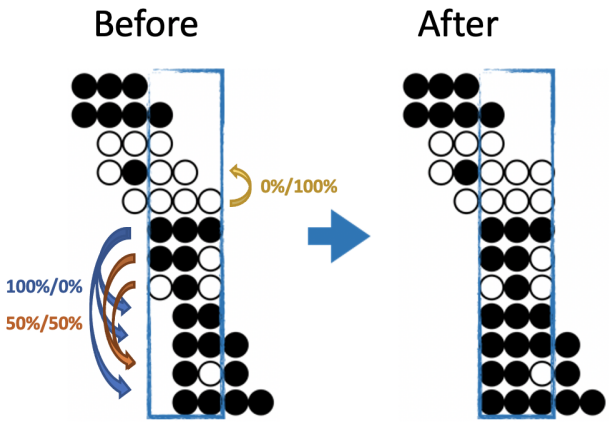

Figure 1: An illustration of both the eligibility and a possible imputation result for BSImp. Given we are interested in window size of three cytosines, a region is selected as enclosed by a blue rectangle. Each line of dots represents a read aligned to a specific genomic region; black (white) dots represents a methylated (unmethylated) cytosine for the given read. Four complete patterns are observed, which makes the window eligible for imputation. Within the window, only reads missing at most one methylation statuses (dots) are eligible for imputation; there are five in the example. Looking at the topmost reads with one missing pattern, the rest of the pattern resembles that of the complete pattern below, so it has a probability of zero being methylated. The second last read has methylation pattern (black, white in the second and third position) resembles two other reads, which have methylation statuses of methylated and unmethylated, one each, so it has 50% of being either.

In our implementation, the imputations are done alongside genome screening where windows of fixed size of cytosines of the same methylation contexts are extracted, imputed if valid and profiled for their copy numbers of methylated, unmethylated reads and every possible methylation patterns. It is done through sliding windows with one cytosine overlapping. Only windows with at least two complete patterns are considered for imputation and results

outputted if a given cytosine has enough depths as specified by the user. A visualisation of the process can be found in Figure 2 and an example output in Figure 3.

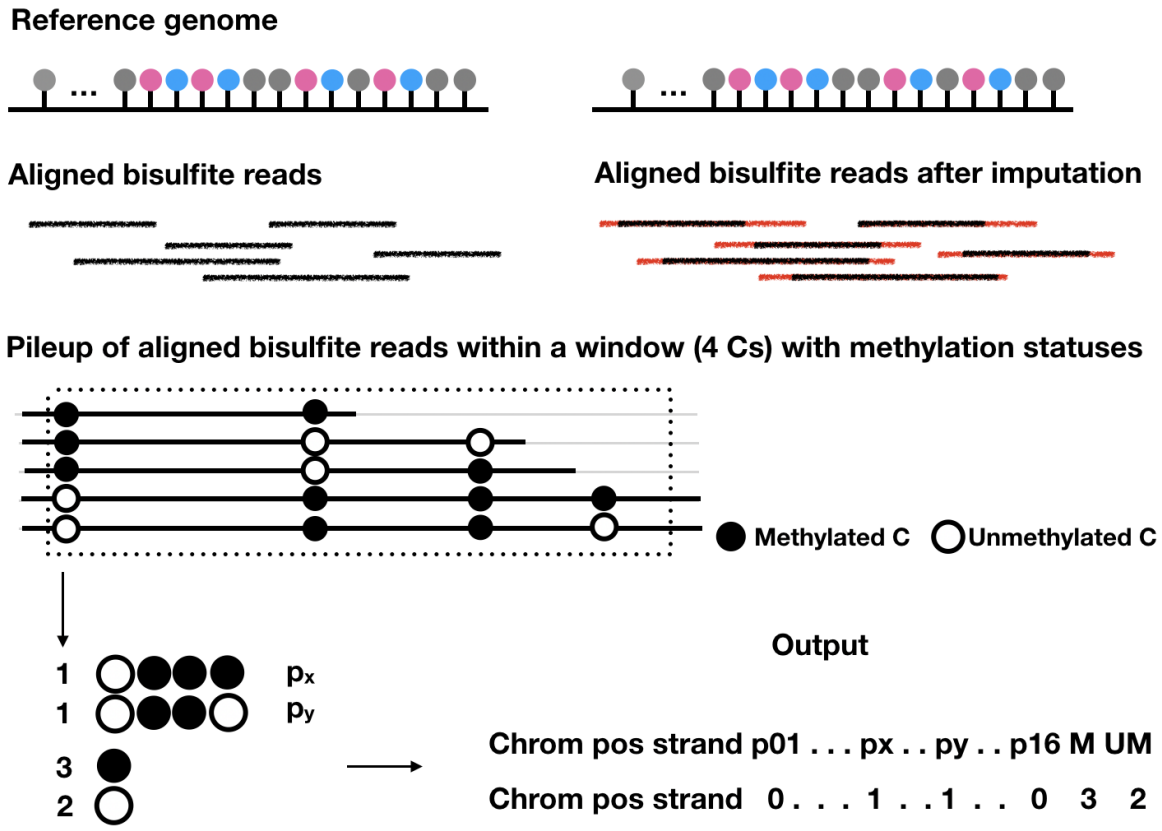

Figure 2: Schematic representation of the output of BSImp. First the reads are aligned to the reference genome, given any window of specified number of cytosines (here is 4), methylation statuses are extracted and missing values imputed if valid, then methylation profiled as a vector including chromosome, position, strand, the copy numbers of all possible methylation patterns starting at the position, and number of methylated and unmethylated cytosines.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	chrom	pos	p01	p02	p03	p04	p05	p06	p07	p08	p09	p10	p11	p12	p13	p14	p15	p16	M	UM	strand
2	1	109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	f
3	1	115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	f
4	1	161	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	4	f
5	1	310	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	13	f
6	1	110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	2	r
7	1	116	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	r
8	1	162	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	r
9	1	500	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	7	36	4	f
10	1	511	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	28	52	5	f
11	1	642	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	30	32	2	f
12	1	647	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4	34	0	f
13	1	501	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	4	1	r
14	1	512	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	18	23	1	r
15	1	643	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	19	22	1	r

Figure 3: An example output.

Format desctiptions:

- chrom: chromosome
- pos: (starting cytosine) position for methylation patterns and position for read copy number
- pxx: copy number of methylation pattern
  p01: '0000' - UUUU - copy number of methylation pattern: all unmethylated
  p02: '1000' - MUUU
  p03: '0100' - UMUU
  p04: '1100' - MMUU
  p05: '0010' - UUMU
  p06: '1010' - MUMU
  p07: '0110' - UMMU
  p08: '1110' - MMMU
  p09: '0001' - UUUM
  p10: '1001' - MUUM
  p11: '0101' - UMUM
  p12: '1101' - MMUM
  p13: '0011' - UUMM
  p14: '1011' - MUMM

p15: '0111' - UMMM
p16: '1111' - MMMM - copy number of methylation pattern: all methylated

- M: # of methylated C/G
- UM: # of unmethylated C/G (T/A)
- strand: f(orward)/r(everse)