A New Deployment of the Burrows-Wheeler Transform to Improve DNA Methylation Analysis

Ricardo Olanda Rodríguez
Mariano Pérez Martínez
Juan M. Orduña Huertas
César González Segura

Dpto. de Informática, Universidad de Valencia
Juan.orduna@uv.es

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD

Escola Tècnica Superior d'Enginyeria

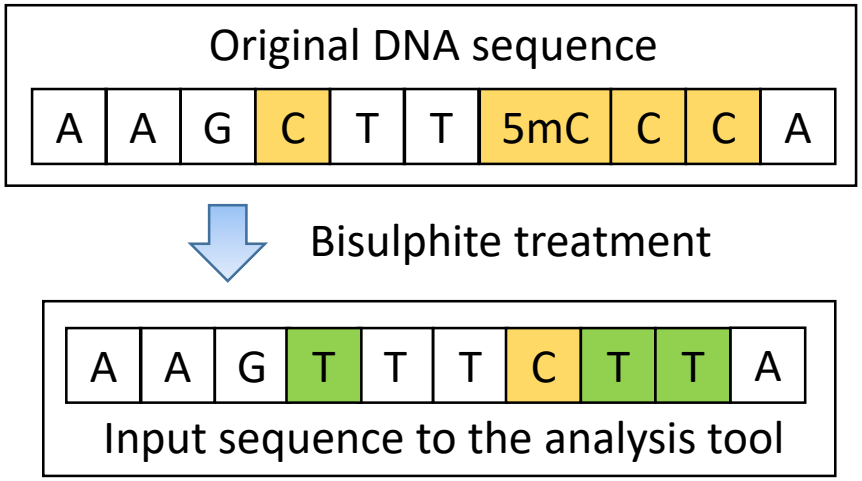
Introduction

- The introduction of Next Generation Sequencing (NGS) has increased the size and length of DNA samples.
- Current sequencers produce billions of short DNA samples, with lengths usually surpassing 400 nt[1].
- Upcoming sequencers will be able to generate samples with lengths up to thousands of nts, producing huge datasets as a result [2].

* [Illumina NextSeq 500 (https://research.ncsu.edu/gsl/files/2015/08/Illumina_NextSeq_500.jpg)]

DNA Methylation

- Methylation is a particular topic of DNA analysis, relevant in the study of cancer, diabetes and other diseases.
- Through bisulfite-treatment, the methylation status can be retrieved with base-pair resolution.



- Several methylation analysis tools have been developed[3,4,5,6], however their performance decreases with read lengths over 150 nts.

Methylation Analysis Pipeline

- Our approach is based on using two mapping algorithms, the Burrows Wheeler Transform[7] and the Smith & Waterman Algorithm[8].

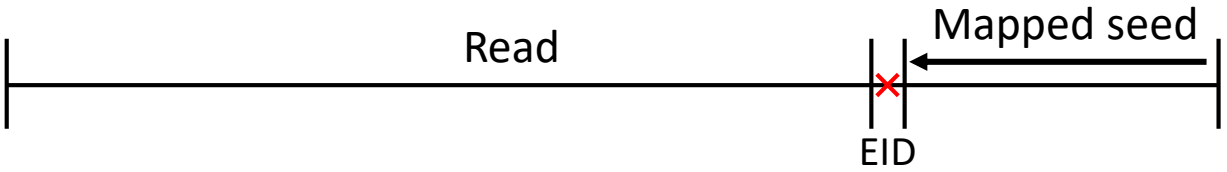
- In HPG-Methyl 2, we propose a new strategy to reduce the execution time and increase sensitivity even for long reads using two techniques:

  - A bidirectional BWT, which can discard wrong alignments more efficiently.

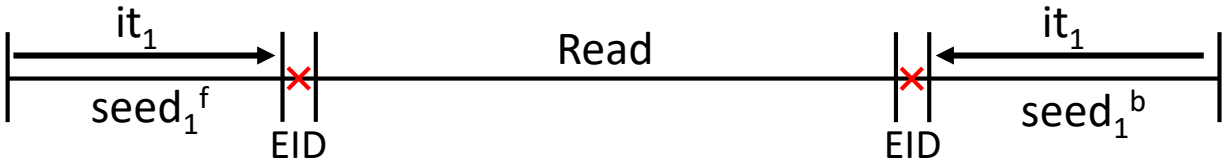
  - A new parallel pipeline scheme, reducing the number of invalid candidate regions and the execution time of the SWA.

New Implementation of the Burrows-Wheeler Transform

- The Burrows-Wheeler transform can be used as a backward search method, allowing errors, insertions or deletions (EIDs)[7].
- Unidirectional BWT performs a backward search, mapping a read segment to the reference genome base by base, until an EID is found.



- Bidirectional BWT performs the search from both ends simultaneously, doubling the maximum number of allowed EIDs.

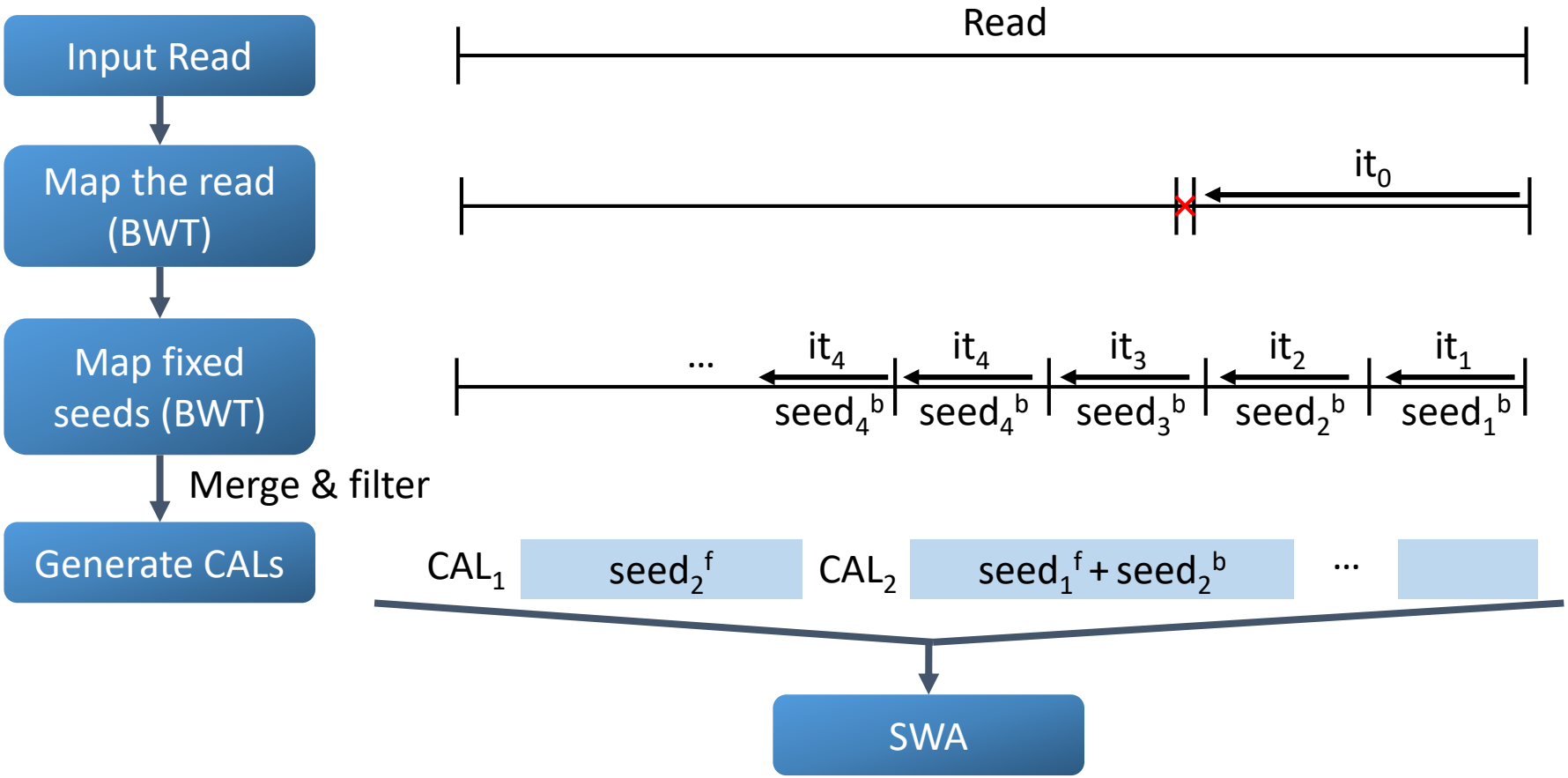


Implementation in a New Parallel Pipeline

- We have used HPG-Methyl[9], a parallel methylation analysis tool, as the basis to implement these new strategies.
- To take full advantage of the new BWT implementation, HPG-Methyl features several changes to the pipeline:
  - Several stages are merged into a single, more flexible stage.
  - This new stage produces fewer candidate regions (CALs), but much more effective.
  - As a result, use of the SWA algorithm is greatly reduced, improving execution times.

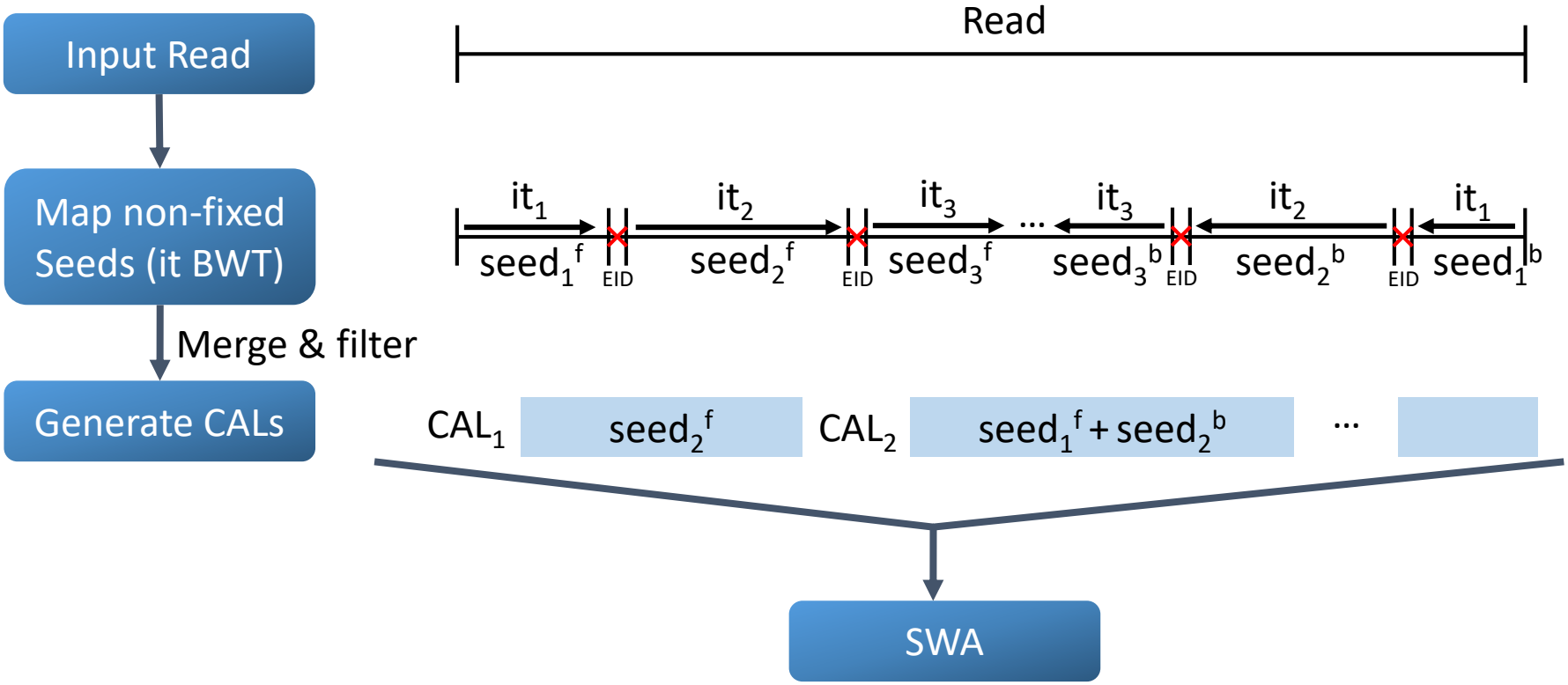
Implementation in a New Parallel Pipeline

- Stages of the original pipeline in HPG-Methyl.



Implementation in a New Parallel Pipeline

- Modified stages of the parallel pipeline in HPG-Methyl 2.



Performance Evaluation

The performance of HPG-Methyl 2 was compared with HPG-Methyl[9] and Bismark[3], using:

- Synthetic methylation datasets with read lengths between 75 nt and 3200 nt, with 4 million reads.
- Real methylation datasets from the European Nucleotide Archive[10].

Tests were conducted in a computer platform as in the table, measuring:

- Sensitivity.
- Execution time.

Computer Platform
Intel i7 3930K (6 cores @ 3,8 GHz)
48 GB RAM
SSD Storage
Ubuntu 14.04 LTS

Performance Evaluation: sensitivity

- HPG-Methyl 2 yields the most accurate results, with sensitivity rates over 99% in all cases.

Length (nt)	HPG-Methyl 2		HPG-Methyl		Bismark	
	Right	Wrong	Right	Wrong	Right	Wrong
75	**99,50%**	0,69%	93,37%	0,62%	88,30%	0,10%
150	**99,01%**	0,46%	96,87%	0,80%	94,59%	0,08%
400	**99,75%**	0,18%	97,55%	0,48%	97,55%	0,10%
800	**99,93%**	0,06%	96,94%	0,48%	98,45%	0,08%
1600	**99,75%**	0,06%	96,94%	0,48%	—*	—*
3200	**99,68%**	0,08%	96,42%	0,49%	—*	—*

* Aborted after waiting for 3 days (4320 minutes)

Performance Evaluation: execution time

- Our new implementation allows for a significant reduction in the execution time.
- Now HPG-Methyl scales with the increasing sequence lengths of NGS.

Length (nt)	HPG-Methyl 2	HPG-Methyl	Bismark
75	**1,288**	1,366	69,579
150	**1,550**	1,950	106,173
400	**5,041**	10,85	248,107
800	**11,26**	50,60	1246,89
1600	**34,44**	996,567	—*
3200	**164,593**	7733,38	—*

* Aborted after waiting for 3 days (4320 minutes)

Conclusions

- We have proposed a new strategy for methylation analysis, using a bidirectional BWT implementation and a new parallel pipeline.
- Our implementation reduces execution times by an order of magnitude and achieves sensitivity rates over 99%, even for very long reads.
- The software is open source and available in Github: `https://github.com/grev-uv/hpg-methyl`

Future Work

- This strategy can be exported to other methylation analysis tools, and to DNA and RNA analysis tools.
- We plan on applying the same strategy of bidirectional BWT to similar tools, such as HPG-Aligner.
- Also, HPG-Methyl will keep being updated with performance improvements and new features.

References

[1] Inc. PACBIO Pacific Biosciences of California, “PACBIO RS II: The original long-read sequencer,” http://www.pacb.com/products-and-services/pacbio-systems/rsii/, 2016.

[2] Joaquin Tarraga, Asunción Gallego, Vicente Arnau, Ignacio Medina, and Joaquin Dopazo, “Hpg pore: an efficient and scalable framework for nanopore sequencing data.,” BMC bioinformatics, vol. 17, pp. 107, 2016 2016.

[3] Felix Krueger and Simon R. Andrews, “Bismark: a flexible aligner and methylation caller for bisulfite-seq applications,” Bioinformatics, vol. 27, no. 11, pp. 1571–1572, 2011.

[4] Pao-Yang Chen, Shawn Cokus, and Matteo Pellegrini, “Bs seeker: precise mapping for bisulfite sequencing.,” BMC Bioinformatics, vol. 11, pp. 203, 2010.

[5] Yuanxin Xi and Wei Li, “BSMAP: whole genome bisulfite sequence MAPping program.,” BMC bioinformatics, vol.10, no. 1, pp. 232+, 2009.

[6] Yuanxin Xi, Christoph Bock, Fabian Maller, Deqiang Sun, Alexander Meissner, and Wei Li, “Rrbsmap: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing.,” Bioinformatics, vol. 28, no. 3, pp. 430–432, 2012.

[7] H. Li and R. Durbin, “Fast and accurate short read alignment with burrows-wheeler transform.,” Bioinformatics, vol. 25, no. 14, pp. 1754–60, 2009.

[8] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences.,” Journal of molecular biology, vol. 147, no. 1, pp. 195–197, Mar. 1981.

[9] Joaquín Tárraga, Mariano Pérez, Juan M Orduña, José Duato, Ignacio Medina, and Joaquín Dopazo, “A parallel and sensitive software tool for methylation analysis on multicore platforms.,” Bioinformatics (Oxford, England), vol. 31, no. 19, pp. 3130–3138, 2015 Jun 10 2015.

[10] http://www.ebi.ac.uk/ena/data/view/SRR309230 and http://www.ebi.ac.uk/ena/data/view/SRR837425

A New Deployment of the Burrows-Wheeler Transform to Improve DNA Methylation Analysis

Ricardo Olanda Rodríguez
Mariano Pérez Martínez
Juan M. Orduña Huertas
César González Segura

Dpto. de Informática, Universidad de Valencia
Juan.orduna@uv.es

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD
AGENCIA ESTATAL DE INVESTIGACIÓN

Escola Tècnica Superior d'Enginyeria