BAllC and BAllCools: Efficient Formatting and Operating for Single-Cell DNA Methylation Data

Wei Tian[1], Wubin Ding[1] and Joseph R. Ecker[1,2,*]

[1]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA, [2]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA.

*To whom correspondence should be addressed.

Abstract

Motivation: With single-cell DNA methylation studies yielding vast datasets, existing data formats struggle with the unique challenges of storage and efficient operations, highlighting a need for improved solutions.
Results: BAllC (Binary All Cytosines) emerges as a tailored binary format for methylation data, addressing these challenges. BAllCools, its complementary software toolkit, enhances parsing, indexing, and querying capabilities, promising superior operational speeds and reduced storage needs.
Availability: https://github.com/jksr/ballcools
Contact: ecker@salk.edu

1 Introduction

The advent of single-cell DNA methylation research has significantly enhanced our grasp of epigenetic intricacies, bringing to light the heterogeneities existing across individual cells – heterogeneities that are often masked in bulk assays (Luo *et al.*, 2017; Liu *et al.*, 2021; Tian *et al.*, 2022). However, the granularity and scope of such investigations yield immense datasets, often spanning tens to hundreds of terabytes, presenting storage and processing challenges. While various specialized formats have been developed for DNA methylation profiling (Argelaguet *et al.*, 2019; Li *et al.*, 2021; Ding *et al.*, 2022; Loyfer *et al.*, 2023), many fail to adequately address the distinct sparsity of single-cell DNA methylation, encompass both CpG and non-CpG methylation contexts, or achieve the optimal operational speed and storage efficiency. Recognizing the pressing need for a more efficient alternative, we embarked on the development of BAllC (Binary All Cytosines), a tailored binary format for DNA methylation data. Accompanied by its software toolkit, BAllCools, this system aims to counter the limitations of existing formats, offering researchers a streamlined approach for the next era of epigenetic exploration.

2 Methods

2.1 The BAllC format

The BAllC format comprises a header and a methylation record section (Figure 1). The header identifies the BAllC version, provides genome reference information, and accommodates user notes. The methylation record section contains genome coordinates and both methylated and total counts for each cytosine. A supplementary TAB-delimited CMeta file can store positions and contexts of each cytosine in the genome (Table S1). Only one CMeta file is needed for multiple BAllC files with the same genome reference, reducing redundancy in data storage. Both BAllC and CMeta files use the blocked compressed format (BGZF (Bonfield *et al.*, 2021)) and can be indexed due to stored genome coordinates, enabling efficient data access.

Field	Description	Type	Value
header			
magic	magic string	char[6]	"BALLC\1"
version	major version of the BAllC format	uint8_t	e.g. "0"
version_minor	minor version of the BAllC format	uint8_t	e.g. "1"
sc	flag of single cell BAllC format	uint8_t	"0" or "1"
l_assembly	length of assembly text	uint32_t	
assembly_text	user specified assembly info	char[l_assembly]	e.g. "mm10", "hg38", "donor1"
l_text	length of header text	uint32_t	
header_text	note for this ballc file	char[l_text]	
n_refs	number of references (chromsomes)	uint16_t	
refs	list of refs (n=n_refs)		
ref			
l_name	length of reference name	uint32_t	
ref_name	reference name	char[l_name]	e.g. "chr1"
ref_len	length of reference	uint32_t	e.g. "248956422"
mc_records	list of mc_records		
mc_record			
pos	position of cytosine	uint32_t	
ref_id	reference id, 0 ≤ref_id < n_refs	uint16_t	
mc	number of counts supporting methylation call	uint16_t when sc=0; uint8_t when sc=1	
cov	number of counts covering the position	uint16_t when sc=0; uint8_t when sc=1	

Table. 1. Specification of the BAllC format.

2.2 BAllCools software package

BAllCools is a C/C++ library and software package designed for BAllC format operations. It supports format conversion between BAllC and the AllC text-based format (Schultz *et al.*, 2015), provides indexing and querying capabilities for BAllC files, and offers efficient methylation information retrieval based on genome range or cytosine context. Notably, BAllCools can merge multiple BAllC files rapidly. A Python wrapper is also available for flexibility.

3 Results

We evaluated the efficiency of BAllC/BAllCools using large public single-cell DNA methylation datasets (Liu *et al.*, 2021; Tian *et al.*, 2022) and benchmarked them against the text-based AllC format (Schultz *et*

al., 2015) and the associated toolkit AllCools (Liu *et al.*, 2021), both of which are extensively used in studies of single cell DNA methylation.

3.1 Storage efficiency

Due to its binary nature and minimized redundancy, the BAllC format significantly optimizes storage in comparison to text-based formats. BAllC achieved an impressive >52% storage reduction when compared to the compressed (gzipped) text-based format, and a >93% reduction in comparison to the uncompressed format.

3.2 Operational speed

Merging multiple single-cell DNA methylomes into pseudo-bulk data is one of the most frequent operations in processing single-cell DNA methylation datasets, facilitating further analyses like differential methylation calling and visualization on genome browsers. Merging with BAllCools drastically cuts down the time required in comparison to text-based alternatives (Figure S1). For instance, merging a typical set of 500 single cell DNA methylomes takes only about 2 hours using one CPU core (2450MHz) , resulting in a time savings of over 30x.

3.3 Flexible data query

Beyond allowing queries of DNA methylation records for specific genome coordinates, BAllCools provides the flexibility to filter results by cytosine context. Users can retrieve methylation statuses of only CpG sites or cytosines within contexts of particular interest of users, such as CTG (Figure S2). Such level of flexibility is often unattainable with the text-based format and its associated tools..

4 Conclusions

With the increasing volume of single-cell DNA methylation data, efficient storage and processing solutions like BAllC and BAllCools are paramount. Our binary format, coupled with a comprehensive toolkit, presents a notable advancement in handling such datasets, ensuring optimal performance in storage, speed, and data retrieval.

Acknowledgements

We thank Jingtian Zhou for assistance in testing and suggestive comments. We are very grateful to members of the Ecker group for their feedback and discussion.

Funding

This work was supported by grants from NIMH U19MH11483, U01MH121282 and UM1MH130994 to J.R.E.

Conflict of Interest: J.R.E is a member of the scientific advisor for Zymo Research In. and Ionis Pharmaceuticals.

References

Argelaguet,R. *et al.* (2019) Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, **576**, 487–491.

Bonfield,J.K. *et al.* (2021) HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience*, **10**.

Ding,Y. *et al.* (2022) mHapTk: a comprehensive toolkit for the analysis of DNA methylation haplotypes. *Bioinformatics*, **38**, 5141–5143.

Li,L. *et al.* (2021) Dissecting the epigenomic dynamics of human fetal germ cell development at single-cell resolution. *Cell Res.*, **31**, 463–477.

Liu,H. *et al.* (2021) DNA methylation atlas of the mouse brain at single-cell resolution. *Nature*, **598**, 120–128.

Loyfer,N. *et al.* (2023) A DNA methylation atlas of normal human cell types. *Nature*, **613**, 355–364.

Luo,C. *et al.* (2017) Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, **357**, 600–604.

Schultz,M.D. *et al.* (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, **523**, 212–216.

Tian,W. *et al.* (2022) Epigenomic complexity of the human brain revealed by single-cell DNA methylomes and 3D genome structures. *bioRxiv*, 2022.11.30.518285.