*Bioinformatics*, 37(18), 2021, 3011–3013
doi: 10.1093/bioinformatics/btab148
Advance Access Publication Date: 3 March 2021
Applications Note

Genome analysis

# S3V2-IDEAS: a package for normalizing, denoising and integrating epigenomic datasets across different cell types

**Guanjue Xiang [1,*], Belinda M. Giardine[2], Shaun Mahony[2], Yu Zhang[3] and Ross C. Hardison [2,*]**

[1]The Bioinformatics and Genomics Program, Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA, [2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA and [3]Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

*To whom correspondence should be addressed.
Associate Editor: Peter Robinson

Received on September 15, 2020; revised on January 26, 2021; editorial decision on February 28, 2021; accepted on March 1, 2021

## Abstract

**Summary:** Epigenetic modifications reflect key aspects of transcriptional regulation, and many epigenomic datasets have been generated under different biological contexts to provide insights into regulatory processes. However, the technical noise in epigenomic datasets and the many dimensions (features) examined make it challenging to effectively extract biologically meaningful inferences from these datasets. We developed a package that reduces noise while normalizing the epigenomic data by a novel normalization method, followed by integrative dimensional reduction by learning and assigning epigenetic states. This package, called S3V2-IDEAS, can be used to identify epigenetic states for multiple features, or identify discretized signal intensity levels and a master peak list across different cell types for a single feature. We illustrate the outputs and performance of S3V2-IDEAS using 137 epigenomics datasets from the VISION project that provides **V**al**I**dated **S**ystematic **I**ntegrati**ON** of epigenomic data in hematopoiesis.
**Availability and implementation:** S3V2-IDEAS pipeline is freely available as open source software released under an MIT license at: https://github.com/guanjue/S3V2_IDEAS_ESMP.
**Contact:** rch8@psu.edu or gzx103@psu.edu
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

The tens of thousands of epigenomic datasets now available are potentially great resources to better understand the associations of epigenetic modifications with mechanisms of transcriptional regulation (Bernstein et al., 2010; ENCODE Project Consortium, 2012; Martens and Stunnenberg, 2013; Moore *et al.*, 2020; Stunnenberg *et al.*, 2016; Xiang, et al., 2020a, b; Yue *et al.*, 2014). However, integrating these resources for global inferences about regulation is challenging for many reasons. In this application note, we focus on two issues. First, technical differences in procedures and biological samples analyzed in different laboratories introduce noise and biases that can obscure true biological differences (Meyer and Liu, 2014; Shao et al., 2012; Xiang et al., 2020a,b). Second, certain combinations of epigenetic modifications often appear together, but those combinations of modifications (epigenetic states) need to be learned from integrative modeling across epigenomic datasets simultaneously across multiple cell types (Ernst and Kellis, 2012; Hoffman et al., 2012; Zhang et al., 2016).

Here, we introduce a package, named S3V2-IDEAS, that builds upon our prior works and provides an improved, integrated workflow that will facilitate usability. In this pipeline, we address the first issue (noise and bias in data) by incorporating an improved version of the S3norm method (Xiang et al., 2020b), which can simultaneously normalize signals in foreground and signals in background. In contrast to S3norm, in which each 200 bp bin was assessed as either foreground (peak) or background, in the improved version (S3V2) the reads within each bin are split into foreground reads and background reads. This strategy has been used in several previous studies (Mahony et al., 2014; Tarbell and Liu, 2019). After splitting reads, a single signal track can be converted into a foreground signal track and a background noise track. For the background noise track, both non-zero mean and non-zero standard deviations are matched across datasets, which can reduce the noise in some datasets (Fig. 1A, Supplementary Methods and Supplementary Figs S1D and S2). To address the second challenge (integration across multiple features and cell types), we performed genome segmentation using

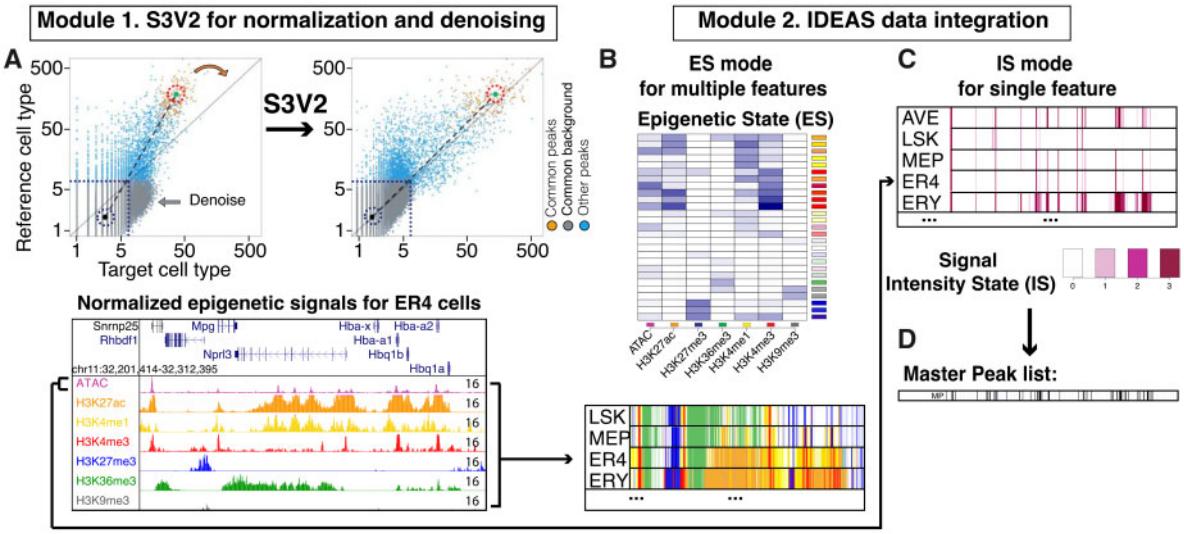


**Fig. 1.** Overview of S3V2-IDEAS pipeline. (**A**) Module 1 normalizes and denoises input data using the S3V2 method. Examples of normalized epigenetic signals from the *Hba* locus in G1E-ER4 cells (ER4) are shown. (**B** and **C**) In Module 2, the normalized data is integrated by IDEAS in one of two modes. (B) The epigenetic state mode integrates multiple epigenetic features into an epigenetic states model. (C) The signal intensity state mode finds frequently occurring signal intensity states for a single epigenetic feature, along with a master peak list derived from those states (**D**). AVE = average, LSK, MEP, ER4, ERY = abbreviations for cell types (Xiang et al., 2020a,b)

the Integrative and Discriminative Epigenome Annotation System (IDEAS), which learns epigenetic state models from the normalized epigenomic signals simultaneously along the genome and across cell types to improve consistency of state assignments across different cell types (Zhang *et al.*, 2016; Zhang and Hardison, 2017). Moreover, the IDEAS model can jointly estimate the state of a genomic region by using the information in a set of similar cell types, so that the state can be accurately estimated even for cell types with missing data (Zhang and Mahony, 2019). The S3V2-IDEAS pipeline incorporates both S3V2 normalization and IDEAS segmentation so that the advantages of both methods can be used to normalize, denoise and integrate multi-dimensional epigenomic datasets across different cell types.

## 2 Implementations

The inputs to S3V2-IDEAS are (i) average read counts of each epigenetic feature in each cell type (bigWig), (ii) an annotation file that includes the names of the cell types and the epigenetic features of bigwig files and (iii) information about the mapped genome, such as chromosome sizes and black-listed regions (Amemiya et al., 2019; Boyle et al., 2014; Kent et al., 2002, 2010; Yue *et al.*, 2014).

The S3V2-IDEAS incorporates two major modules. First, it uses the S3V2 method to normalize and denoise the epigenomic datasets (Fig. 1A). The second module of the package incorporates the IDEAS genome segmentation model to integrate the epigenomic signal into tracks of epigenetic state assignments for each bin in each cell type (Fig. 1B and C). The second module can operate in either of two modes. When the input data include multiple epigenetic features, the module executes an epigenetic states mode (ES mode), which integrates the signals of multiple epigenomic features into epigenetic states as done previously (Fig. 1B). When the input data include one epigenetic feature, the module executes a signal intensity state mode (IS mode) to quantize the signal of that one epigenomic feature into discrete signal levels (Fig. 1C). In the IS mode, a master peak list (Fig. 1D) can be extracted from the signal intensity state tracks by a novel hierarchical method which provides way to integrate the epigenomic information across cell types (Supplementary Figs S5 and S6).

## 3 Results and discussion

The S3V2-IDEAS produces three outputs: the normalized signal tracks and the -log10 *P*-value tracks based on the background model (Fig. 1A); a list of epigenetic states or signal intensity states and the corresponding state track in each cell type (Fig. 1B and C). An additional master peak list can be produced in the IS mode (Fig. 1D).

To illustrate these results, we applied the S3V2-IDEAS to datasets compiled by the ValIdated and Systematic integratION of epigenomic data project (VISION) (Hardison et al., 2020; Heuston *et al.*, 2018; Xiang et al., 2020a,b). The ES mode can integrate seven epigenetic features to a 27 epigenetic states model (Fig. 1B). Compared with our previous analysis (Xiang et al. 2020a,b), the genome segmentation tracks from this S3V2-IDEAS are more consistent between biological replicates (Supplementary Figs S2C, D and S3D). The de-noising by S3V2 improves the accuracy of peak calling in ChIP-seq datasets (Supplementary Fig. S2E).

We illustrate the IS mode by limiting our analysis to only the ATAC-seq. In the IS mode, the ATAC-seq signal tracks can be first normalized and discretized into tracks of signal intensity levels (Fig. 1C). Then, a master peak list can be extracted from these state tracks (Fig. 1D). A master peak list is a straightforward way to obtain a coherent set of chromatin accessible peaks across cell types, which can be challenging for larger numbers of cell types (Meuleman et al., 2020). Comparing with the one produced by simply pooling and merging the MACS2 peaks in all cell type (Zhang et al., 2008), the IS mode master peak list pinpoints functional elements with higher accuracy (Supplementary Figs S6 and S7).

These results indicate that the S3V2-IDEAS should be versatile and effective tool for integrative analyses of epigenomic datasets.

## Funding

The work was supported by National Institutes of Health [R01 GM121613 and R24 DK106766].

*Conflict of Interest:* none declared.

## References

Amemiya,H.M. *et al.* (2019) The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.*, **9**, 9354.

Bernstein,B.E. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.

Boyle,A.P. *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**, 453–456.

ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Hardison,R.C. *et al.* (2020) Systematic integration of GATA transcription factors and epigenomes via IDEAS paints the regulatory landscape of hematopoietic cells. *IUBMB Life*, **72**, 27–38.

Heuston,E.F. *et al.*; NIH Intramural Sequencing Center. (2018) Establishment of regulatory elements during erythro-megakaryopoiesis identifies hematopoietic lineage-commitment points. *Epigenet. Chromatin*, **11**, 22.

Hoffman,M.M. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.

Kent,W.J. *et al.* (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Mahony,S. *et al.* (2014) An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. *PLoS Comput. Biol.*, **10**, e1003501.

Martens,J.H.A. and Stunnenberg,H.G. (2013) BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, **98**, 1487–1489.

Meuleman,W. *et al.* (2020) Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, **584**, 244–251.

Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, **15**, 709–721.

Moore,J.E. *et al.*; ENCODE Project Consortium. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.

Shao,Z. *et al.* (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.*, **13**, R16.

Stunnenberg,H.G. *et al.*; International Human Epigenome Consortium. (2016) The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell*, **167**, 1145–1149.

Tarbell,E.D. and Liu,T. (2019) HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Res.*, **47**, e91.

Xiang,G. *et al.* (2020a) An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis. *Genome Res.*, **30**, 472–484.

Xiang,G. *et al.* (2020b) S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. *Nucleic Acids Res.*, **48**, e43.

Yue,F. *et al.*; Mouse ENCODE Consortium. (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**, 355–364.

Zhang,Y. and Hardison,R.C. (2017) Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res.*, **45**, 9823–9836.

Zhang,Y. *et al.* (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.*, **44**, 6721–6731.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

Zhang,Y. and Mahony,S. (2019) Direct prediction of regulatory elements from partial data without imputation. *PLoS Comput. Biol.*, **15**, e1007399.