An Integrated CUT&RUN Quality Control Workflow for Histone Modifications and Transcription Factors

Joseph Boyd
University of Vermont
Princess Rodriguez
University of Vermont
Hilde Schjerven
University of California San Francisco
Seth Frietze (seth.frietze@med.uvm.edu)
University of Vermont https://orcid.org/0000-0003-4058-3661

Research note

Keywords: CUT&RUN, ChIP-seq, data quality control, data visualization

Posted Date: June 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-646006/v1

Abstract

Objective

Among the different methods to profile the genome-wide patterns of transcription factor binding and histone modifications in cells and tissues, CUT&RUN has emerged as a more efficient approach that allows for a higher signal-to-noise ratio using fewer number of cells compared to ChIP-seq. The results from CUT&RUN and other related sequence enrichment assays requires comprehensive quality control (QC) and comparative analysis of data quality across replicates. While several computational tools currently exist for read mapping and analysis, a systematic reporting of data quality is lacking. Our aims were to 1) compare methods for using frozen versus fresh cells for CUT&RUN and 2) to develop an easy-to-use pipeline for assessing data quality.

Results

We compared a workflow for CUT&RUN with fresh and frozen samples, and present an R package called ssvQC for quality control and comparison of data quality derived from CUT&RUN and other enrichment-based sequence data. Using ssvQC, we evaluate results from different CUT&RUN protocols for transcription factors and histone modifications from fresh and frozen tissue samples. Overall, this process facilitates evaluation of data quality across datasets and permits inspection of peak calling analysis, replicate analysis of different data types. The package ssvQC is readily available at https://github.com/FrietzeLabUVM/ssvQC.

Introduction

The genome-wide profiling of chromatin-associated proteins and posttranslational modifications (PTMs) to histones has revolutionized research in epigenetic gene regulation [1–3]. One of the mainstay techniques used to map the genome-wide enrichment patterns of PTMs and chromatin-associated proteins is ChIP-seq (Chromatin immunoprecipitation followed by high throughput sequencing), in which proteins are physically crosslinked to target DNA, immunoprecipitated with specific antibodies, and their crosslinks are reversed for downstream deep sequencing and analysis of enriched DNA [4, 5]. ChIP-seq assays typically require a large amount of starting material in the range of millions of cells per immunoprecipitation reaction, and depending on the antibody, can result in a high amount of background. A recently developed method termed Cleavage Under Targets and Release Using Nuclease (CUT&RUN) is becoming widely employed in laboratories for mapping the genomic interactions of proteins [6, 7]. CUT&RUN is an *in situ* genome-wide profiling method that employs an antibody-targeted micrococcal nuclease (MNase) fusion protein to selectively digest and release DNA fragments at protein binding sites. The resulting DNA is used to construct barcoded DNA sequencing libraries that can be pooled and deeply sequenced. Overall, this method results in considerably lower background compared with ChIP-seq [7], and can be employed using much less starting material [8].

While multiple tools exist for CUT&RUN data processing and peak calling [9, 10], there is a need for software to conduct CUT&RUN data quality control (QC), which should be performed prior to detailed data analysis and interpretation. To increase the reproducibility and reliability of the data, a detailed evaluation of the data being produced is essential when performing the assays in different sources of biological material and when testing new antibodies. In the event of a low-quality assay (high background or low signal), a comparative QC process is vital to understanding possible sources of error that may guide investigators to adjust experimental conditions to improve data quality.

Here we introduce ssvQC as a new framework for CUT&RUN data quality assessment. We use this tool to evaluate cell storage protocols for performing CUT&RUN from a limited number of cells and demonstrate that frozen cells from spleen can be used to obtain consistent genome-wide enrichment profiles for transcription factor (TF) and histone modifications generated by CUT&RUN. Overall, ssvQC integrates a set of useful QC metrics to reliably summarize data quality and is implemented as a user-friendly R package.

Material And Methods

Ethics statement

Animal experiments were approved by the University of California San Francisco animal research ethics committee (# AN177290) and the University of Vermont animal research ethics committee (IUACUC# PROTO201900021). All experiments were carried out in accordance with the approved guidelines.

Mice

C57BL/6 mice were obtained from The Jackson Laboratory (Bar Harbor, USA). Mice (6-8 weeks old) of both genders were used for experiments. The animals were bred and maintained either at the animal facilities at University of California San Francisco, or at the University of Vermont. Animals were sacrificed by CO_2 asphyxiation followed by cervical dislocation. A total of four mice were used for all experiments.

Magnetic B cell enrichment

Spleens were harvested postmortem and single-cell suspensions were prepared by gently mashing the spleen through a nylon mesh. Cells were washed once with PBS and red blood cells were lysed with ACK lysis buffer (150mM NH_4Cl, 10mM $KHCO_3$, 0.1mM Na_2EDTA). B cells were isolated using MACS (Miltenyi Biotech, cat # 130-090-862). Purities of enriched B cells of >90% were determined by subsequent flow cytometry analysis. Freshly isolated B cells were then used for either CUT&RUN assays (fresh) immediately or were frozen overnight. Freezing was performed by resuspending cell pellets in 10% DMSO/90% FBS and then slowly frozen overnight at -1°C/minute.

CUT&RUN assays

CUT&RUN was performed on fresh (2 - 8 million cells) or frozen cells (0.5 million cells). CUT&RUN libraries were built using published protocols (EpiCypher, Skene and Henikoff, 2017) with some modifications. Nuclei was isolated from fresh cells using hypotonic lysis buffer (20 mM HEPES-KOH pH 7.9, 10 mM KCl, 1 mM $MgCl_2$, 0.1% Triton X-100, 20% Glycerol) and then washed (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 2 mM EDTA, 0.1% BSA). Frozen cells were thawed for 2 min in a 37°C water bath and washed twice (20mM HEPES, pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1x Roche cOmplete, EDTA-free protease inhibitor). Nuclei or cells were bound with concanavalin A coated magnetic beads (Polysciences) and incubated overnight with primary antibody at 4°C. Unbound antibody was washed away and pAG-MN (EpiCypher) was added and activated with $CaCl_2$ added to a final concentration of 2mM. The reaction was carried out for up to 3 hours at 4°C and quenched with 2X Stop buffer (340mM NaCl, 20mM EDTA, 4mM EGTA, 50 µg/mL RNaseA, 50 µg/mL glycogen). Samples were incubated at 37°C for 20 mins and placed on a magnet to release protein-DNA fragments. Supernatant containing fragments was transferred to a new tube and final volume was raised to 300 µl. To extract DNA, 15 µl of DNA extraction buffer (3 µl of 10% SDS (final concentration 0.1%), 5 µl of proteinase K (at 10 mg/ml), 2 µl RNaseA (at 1mg/ml), and 5 µl of 5M NaCl (final concentration 300mM)) was added, vortexed, and incubated at 50°C for 1 hour. DNA was precipitated by phenol/chloroform/isoamyl alcohol followed by ethanol precipitation with glycogen. DNA pellet was resuspending in TE buffer. Libraries were constructed using the NEBnext Ultra II DNA Library Prep Kit (Illumina) as directed with modifications. To retain fragments > 150 bp after adapter ligation, a 1.1x AMPure XP bead cleanup (Beckman Coulter) was performed. Libraries were PCR amplified for a total of 14 cycles with cycling conditions (1 cycle: 45 sec at 98°C, 14 cycles: 15 sec at 98°C followed by 10 sec 60°C, 1 cycle: 1 min at 72C). The DNA was clean up with 1.1x volume of AMPure XP beads (Beckman Coulter) and fragment size distribution was assessed on DNA Bioanalyzer (Agilent). Samples were pooled at equimolar ratios and sequenced on HiSeq platform using 80 base pair paired end sequencing. The antibodies used in this study were: rabbit anti-H3K4me3 (EpiCypher, cat 13-0041; lot no 20083002-42; 0.5 µg per reaction); rabbit anti-Ikaros (Santa Cruz, sc-13039; 0.5 µg per reaction) and rabbit anti-IgG (Epicypher, cat 13-0042; lot no 20036001-52; 0.5 µg per reaction).

Data processing, peak calling and analysis

For data preprocessing, we used the CUT&RUN tools pipeline as described [10].

Implementation of ssvQC

ssvQC is implemented in R and can run on both MacOS, Windows, and Linux (bigWig operations will not work on Windows). The package can be found at https://github.com/FrietzeLabUVM/ssvQC. The tool is flexible to use and can be run with the minimal inputs of several bam files and corresponding peak files

or can be fully controlled via a plain text configuration file. Data processing operations are implemented in parallel where possible for speed and results are cached to prevent redundant work. All plot outputs are ggplot objects to allow easy customization and intermediate tidy formatted data objects are easily accessible to allow complete control by users. A complete QC report is outputted as a CSV file.

Results

To study the gene regulatory mechanisms of B cells, we aimed to generate genome-wide datasets for regulatory histone modifications and transcription factors (TFs) from a mature B cell population isolated from fresh mouse spleen tissue. Because cell isolation from tissue can be time consuming and the downstream processing of fresh material is not always possible, as in situations with collaborations with other research groups at other locations, we wanted to evaluate if isolated cells could be frozen using a slow freezing protocol (see methods) for CUT&RUN assays and how the results would compare to that of fresh material (**Figure 1A**). We therefore performed a pilot CUT&RUN experiment using antibodies specific for the histone modification H3K4me3, the TF Ikaros and control IgG, using fresh or frozen cells.

To assess the overall data quality, we implemented a uniform data processing pipeline and developed a quality control package called ssvQC to systematically collect and represent CUT&RUN data quality metrics. The overall data processing workflow is illustrated in **Figure 1B**, where each frame represents a preprocessing step leading up to the input data files for ssvQC. The output of preprocessing steps includes mapped BAM files, a signal BigWig file, and the significantly enriched regions in narrowPeak, broadPeak and SEACR bed file format. We also implemented a routine to upload signal files to UCSC genome browser for generating shared sessions, facilitating easy sharing of data with collaborators. ssvQC can similarly be applied to other sequence enrichment assays, including ChIP-seq and ATAC-seq.

We used ssvQC to compare the results of Ikaros and H3K4me3 CUT&RUN profiles generated from fresh and frozen B cell samples. The total number of mapped reads ranged from 6 - 14 million reads per dataset (**S Table 1**). The number of peaks for H3K4me3 was similar between samples, whereas the number of peaks called for Ikaros differed between datasets (10,809 and 22,022 for fresh and frozen samples, respectively). The FRiP score is defined as the fraction of reads that fall into a peak and is often used as a measure of data quality. Each dataset showed a relatively high FRiP scores, > 0.25 for H3K4me3 and > 0.13 for Ikaros (**S Table 1**). Thus, slow freezing of B cells provides equivalent data to that of fresh B cells. Next, we performed CUT&RUN for the transcription factor Ikaros and the histone modification H3K4me3 with frozen biological replicates and compared data quality with ssvQC. Overall, the data was very consistent between replicates for each factor, exhibiting a comparable number of mapped reads, number of peaks and FRiP scores (**Figures 2A-C**). ssvQC was further used to directly compare the overlap of the replicate peak sets for each factor, and depicted a high degree over overlap for each factor across replicates visualized with different graphical outputs (**Figure 2D-F**). The strand cross-correlation (SCC) analysis of ssvQC detects the clustering of reads in CUT&RUN dataset independently of peak calling. ssvQC determines the strand-shift profile based on the position of reads on either the positive or negative strand and shifts them towards one another, and then calculates the correlation

between the position of reads on the two strands at each progressive shift. Overall, each of the datasets produced provided a similar SCC score, providing an assessment of data quality independently of peak call (**Supplemental Figure 1**). Finally, ssvQC can be used to cluster datasets relative to peak regions (or any region) to provide a valuable visual inspection of data quality across different genomic regions of interest (**Figure 3**).

Discussion

Despite the increasing application of CUT&RUN for epigenomic profiling, there is not currently any agreement on experimental protocols starting with different types of biological material, nor is there any routine method for the assessment of data quality. In our experience, fresh biological material from cultured cells or tissues generally provides superior results over frozen material for CUT&RUN. However, when performing lengthy cell isolation procedures such as tissue dissection and cell sorting, and when working with collaborating laboratories at different institutions where shipping of biological material is required, it is not always feasible to proceed directly with CUT&RUN processing. We therefore wanted to evaluate a slow freezing protocol that permits a valuable storage step before performing CUT&RUN. Our analysis indicates that frozen splenic B cells isolated from mice serve as a good starting point for performing CUT&RUN with antibodies against a histone mark and the transcription factor Ikaros.

To compare the quality of datasets from different sample sources, we developed the ssvQC package to incorporate a set of valuable QC metrics that will allow us to evaluate the overall quality of CUT&RUN experiments. The ssvQC package systematically reports these metrics in a clear report with understandable graphics. Because of the datatable implementation, ssvQC can easily process large collections of CUT&RUN, ChIP-seq, ATAC-seq and other sequence enrichment data from both single end and paired end sequencing experiments. Users can reproduce data QC and analysis results in a uniform processing pipeline and integrated software package in R. We further performed replicate frozen CUT&RUN for Ikaros and H3K4me3 from mature splenic B cells and show several features of ssvQC that can be used to assess the quality of replicate datasets. Thus, ssvQC is an effective and versatile solution to systematically process multiple of sequence enrichment datasets.

Limitations

This study employed an analysis of single fresh and frozen samples, and more replicates and conditions, including cell number, could be incorporated to more rigorously evaluate the optimal experimental conditions.

Abbreviations

bp: base pair

SCC: strand cross-correlation

H3K4me3: Histone H3 Lysine 4 trimethyl

FRiP: Fraction of reads in peaks

CUT&RUN: Cleavage Under Targets and Release Using Nuclease

ChIP: Chromatin Immunoprecipitation

MNase: micrococcal nuclease

QC: quality control

RPM: reads per million

Declarations

Ethics approval and consent to participate

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available on the NCBI Gene Expression Omnibus database with accession # GSE172130.

Funding

Research reported in this publication was supported by the National Institutes of Health (NIH) under Award Numbers R01AI127709, R01GM129338, R01CA230618 and U54GM115516. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funding bodies had no role in the design of the study and data collection, analyses and interpretation of data or in the writing of the manuscript.

Acknowledgements

We would like to thank Dr. Eyal Amiel for providing the spleen tissues and Scott Tighe in the Vermont Integrative Genomics Resource for Next Generation Sequencing service and consultation. No financial conflicts of interest exist.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author's Contributions

All authors designed the study. P.R. performed experimental procedures. J.B. undertook the statistical analysis and designed ssvQC. S.F., P.R. and H.S. wrote the manuscript. All authors read and approved the final manuscript.

References

1. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al: **Architecture of the human regulatory network derived from ENCODE data.** *Nature* 2012, **489**:91–100.
2. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**:57–74.
3. Consortium EP: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biol* 2011, **9**:e1001046.
4. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.** *Genome Res* 2012, **22**:1813–1831.
5. O'Geen H, Frietze S, Farnham PJ: **Using ChIP-seq technology to identify targets of zinc finger transcription factors.** *Methods Mol Biol* 2010, **649**:437–455.
6. Meers MP, Bryson TD, Henikoff JG, Henikoff S: **Improved CUT&RUN chromatin profiling tools.** *Elife* 2019, **8**.
7. Skene PJ, Henikoff S: **An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites.** *Elife* 2017, **6**.
8. Skene PJ, Henikoff JG, Henikoff S: **Targeted in situ genome-wide profiling with high efficiency for low cell numbers.** *Nat Protoc* 2018, **13**:1006–1019.
9. Meers MP, Tenenbaum D, Henikoff S: **Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling.** *Epigenetics Chromatin* 2019, **12**:42.
10. Zhu Q, Liu N, Orkin SH, Yuan GC: **CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis.** *Genome Biol* 2019, **20**:192.

Figures

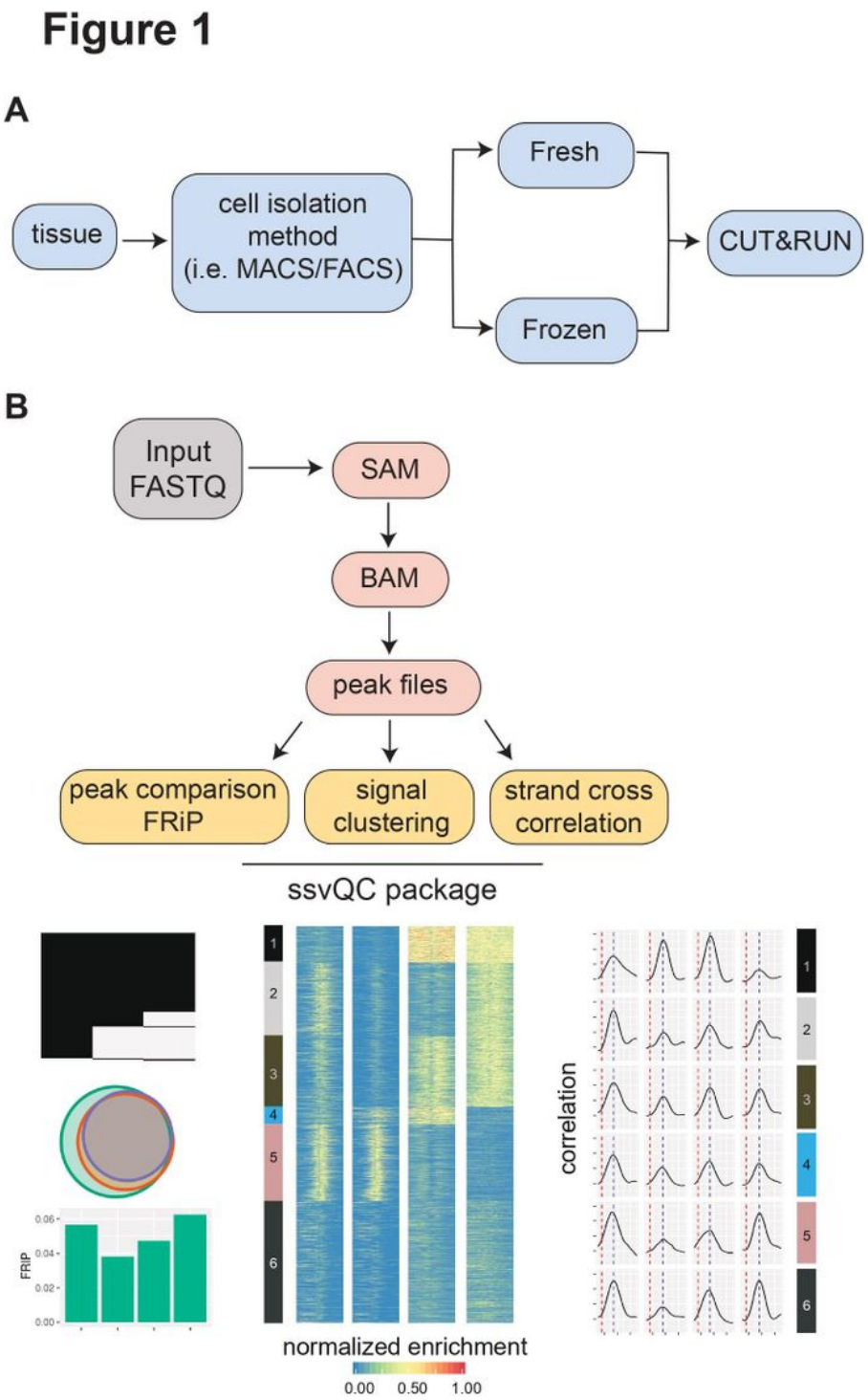

Figure 1

Experimental and data analysis workflows to evaluate CUT&RUN from fresh and frozen samples. A) Experimental process used to compare fresh and frozen cell isolates for CUT&RUN analysis for histone modification and transcription factors. B) Data analysis workflow for ssvQC package. Arrows show the order of the steps. Data preprocessing steps are indicated in salmon color and ssvQC output files are grouped in yellow color. Shown are representative ssvQC output files, including summaries of overlapping

peak sets via binary heatmap or euler plots, clustered heatmap of CUT&RUN signal at peak regions, and strand cross correlation (SCC) analysis showing estimated fragment length (blue line) and read length (red line).

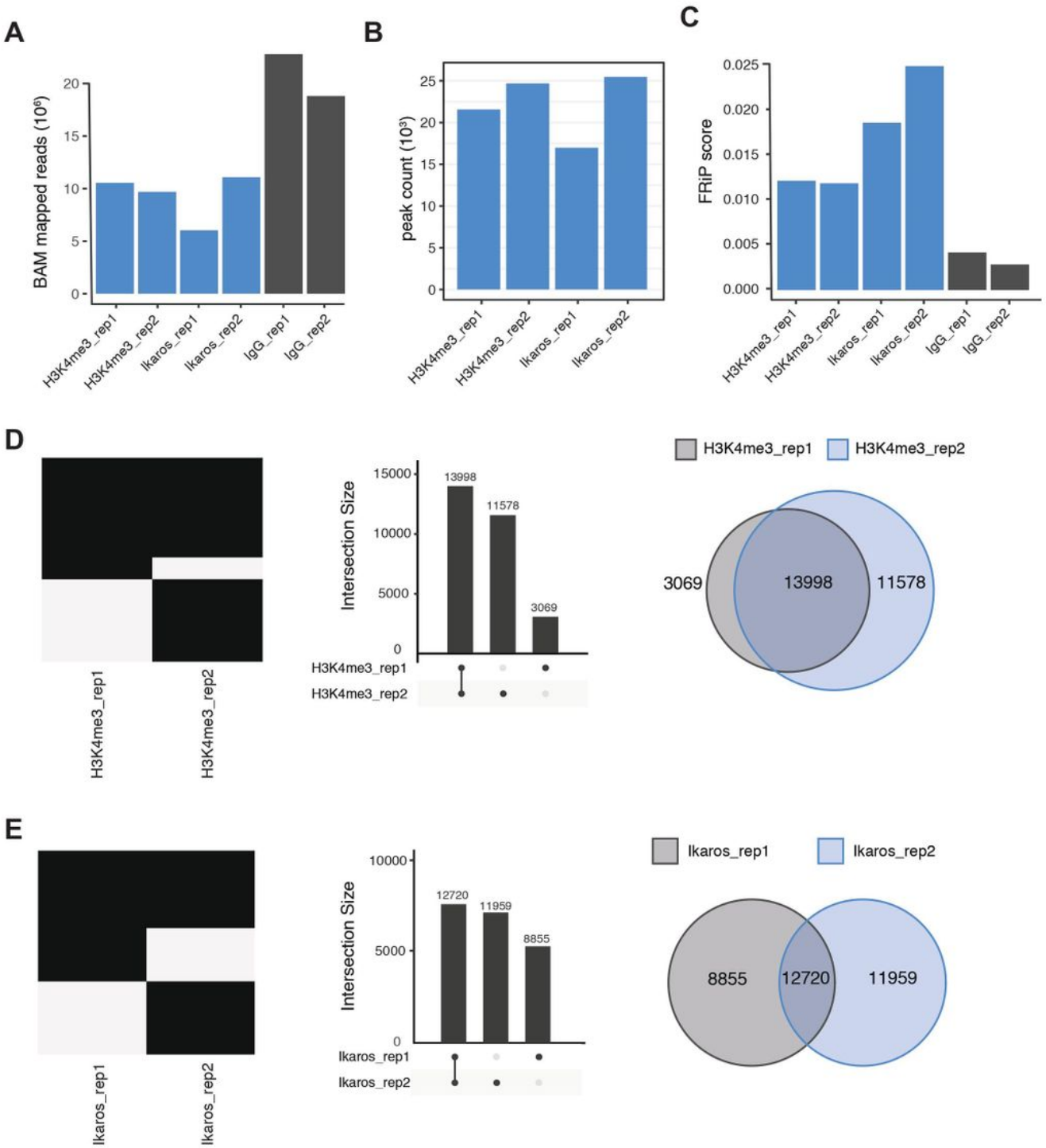

Figure 2

Evaluation of CUT&RUN peaks for transcription factors and histone modifications with replicates using the ssvQC package. The ssvQC output plots for H3K4me3 and Ikaros CUT&RUN replicates from frozen B

cells, showing A) total mapped reads, B) total number of called peaks, C) fraction of reads in peaks (FRiP) scores. The overlaps of peak regions for C) H3K4me3 and D) Ikaros were compared via binary heatmap, upset, and venn diagrams.

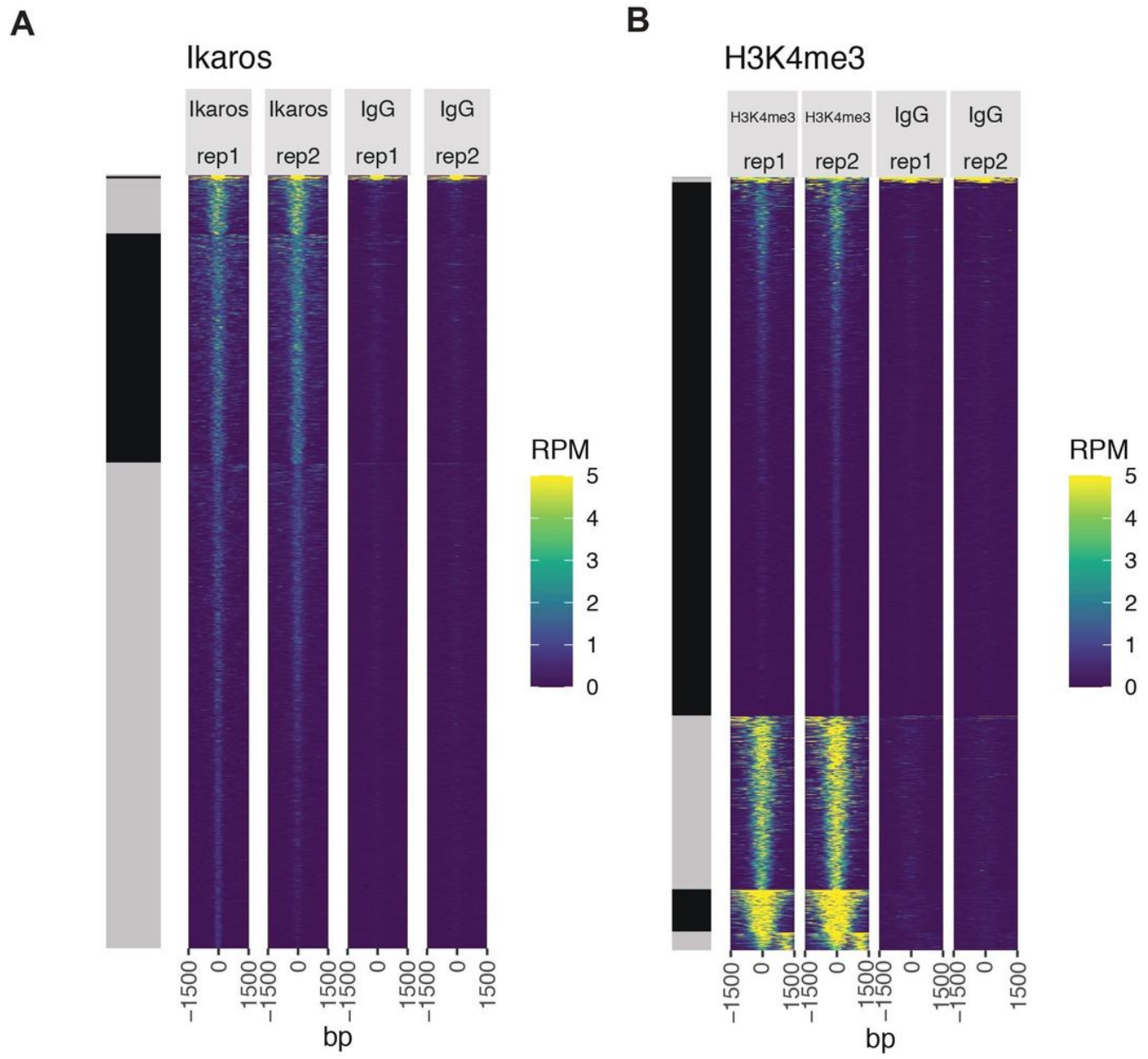

Figure 3

Evaluation of CUT&RUN enrichment profiles with clustered signal heatmap. A) Heatmap showing the normalized (reads per million (RPM) signal of each CUT&RUN dataset over consensus peak regions for A) Ikaros replicate datasets with IgG controls, or B) H3K4me3 replicate datasets with IgG controls. Clustering was performed with kmeans = 6 for both datasets, and shows an overall similar distribution of signal for

called peaks across replicates, exhibiting variable signal across clusters. Also, in either dataset, cluster 1 regions show signal in IgG controls. These peaks can therefore be filtering to obtain a final dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AuthorChecklistcompleted.pdf
- SFig1.pdf
- STable1.xlsx