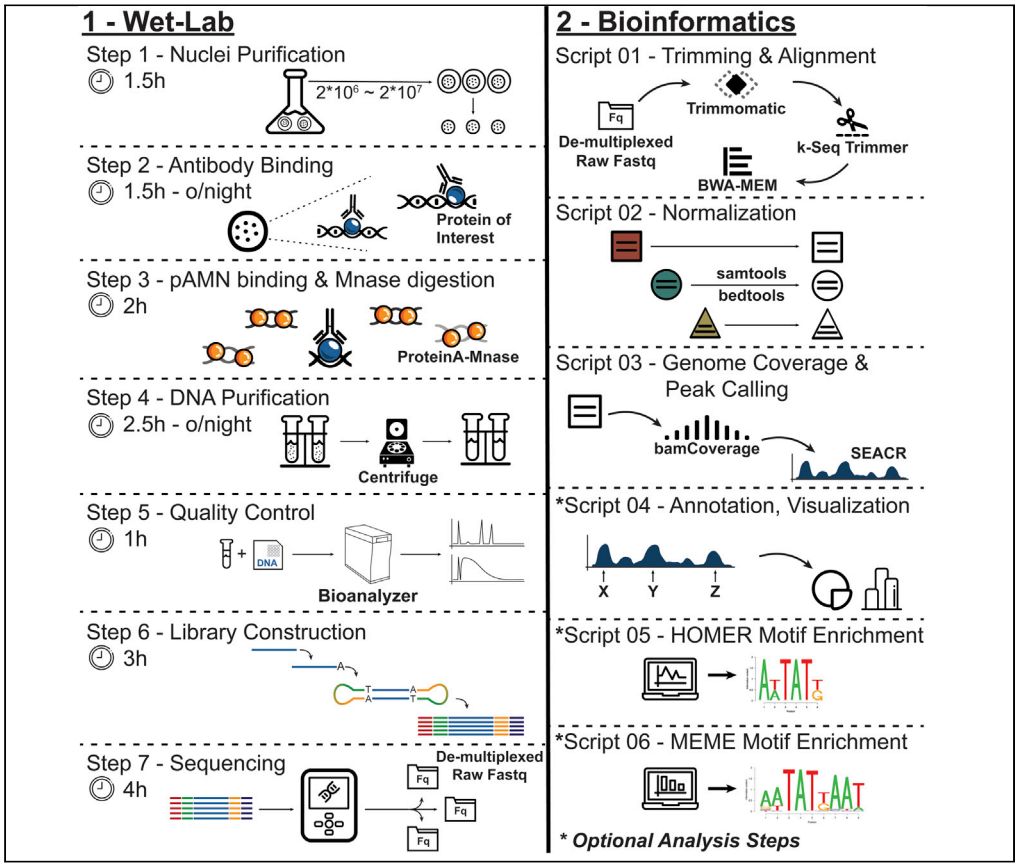
STAR Protocols

CellPress
OPEN ACCESS

Protocol

A modified CUT&RUN protocol and analysis pipeline to identify transcription factor binding sites in human cell lines



CUT&RUN is a recently developed *in situ* chromatin profiling technique that enables high-resolution chromatin mapping and probing. Herein, we describe our adapted CUT&RUN protocol for transcription factors (TFs). Our protocol outlines all necessary steps for TF profiling including the procedure to obtain proteinA-Mnase, while also outlining the bioinformatic pipeline steps required to process, analyze, and identify novel binding sites and sequences. Due to the small number of cells required, this method will allow the elucidation of cell context-dependent functions of many TFs.

Nikki Ruoxi Kong, Li Chai, Daniel Geoffrey Tenen, Mahmoud Adel Bassal

mahmoud.bassal@mymail.unisa.edu.au

Highlights

CUT&RUN was recently developed for *in situ* chromatin mapping and probing

Herein, we describe our modified CUT&RUN protocol to profile TF binding sites and motifs

Modifications relate to nuclear TF targeting, rather than whole-cell histone targeting

Bespoke bioinformatics pipeline simplifies analysis enabling binding site identification

Kong et al., STAR Protocols *2*, 100750
September 17, 2021
https://doi.org/10.1016/j.xpro.2021.100750

STAR Protocols

CellPress
OPEN ACCESS

Protocol

A modified CUT&RUN protocol and analysis pipeline to identify transcription factor binding sites in human cell lines

Nikki Ruoxi Kong,[1,2] Li Chai,[1,2] Daniel Geoffrey Tenen,[2,3] and Mahmoud Adel Bassal[2,3,4,5,*]

[1]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

[2]Harvard Stem Cell Institute, Boston, MA 02115, USA

[3]Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599, Singapore

[4]Lead contact

[5]Technical contact

*Correspondence: mahmoud.bassal@mymail.unisa.edu.au
https://doi.org/10.1016/j.xpro.2021.100750

SUMMARY

CUT&RUN is a recently developed *in situ* chromatin profiling technique that enables high-resolution chromatin mapping and probing. Herein, we describe our adapted CUT&RUN protocol for transcription factors (TFs). Our protocol outlines all necessary steps for TF profiling including the procedure to obtain proteinA-Mnase, while also outlining the bioinformatic pipeline steps required to process, analyze, and identify novel binding sites and sequences. Due to the small number of cells required, this method will allow the elucidation of cell context-dependent functions of many TFs.

For details on the use and execution of this protocol, please refer to Kong et al. (2021).

BEFORE YOU BEGIN

Cleavage under targets and release using nuclease (Cut&Run) is a recently developed method for *in situ* genome wide profiling (Skene et al., 2018). Cut&Run is alternative but analogous technique to chromatin immunoprecipitation with sequencing (ChIP-Seq) that aims to alleviate some of the inherent biases and limitations of ChIP-Seq. In contrast to ChIP-Seq, Cut&Run can be used on as few as 600,000 cells while still showing significant enrichment at transcription factor binding sites (Skene et al., 2018). Furthermore, Cut&Run tends to produce smaller DNA fragments than ChIP-Seq with less background. This translates into requiring shallower sequencing depth and a cleaner, sharper enrichment profile at target sites. This reduced background signal though, wreaks havoc for existing ChIP-Seq peak calling tools and pipelines which require a level of background "noise" for them to adequately call peaks. To that end, a new peak caller was developed, SEACR (Meers et al., 2019), to better handle the low background signal typically seen in Cut&Run datasets and enable peak calling in datasets with sparse background signals.

In this protocol we outline our adapted transcription factor Cut&Run protocol and the bioinformatics pipeline developed to analyze our data. Details of the original study are outlined in our recent publication (Kong et al., 2021). This protocol however, describes in detail the steps undertaken in that study.

The following protocol was performed on SNU-398 cells.

Preparation one: Purification of ProteinA-MNase

Timing: 2 days including overnight (16–18 h) incubation

1. Protein expression
   a. Transform Addgene plasmid number 86973 into BL21-DE3 expression bacteria and plate on kanamycin-containing plates (50 μg/mL), incubate overnight (16–18 h) at 37°C
   b. Next day, pick a colony and grow in 4 mL culture of NZYM media (VWR Catalog number AAJ60918-AP) with kanamycin (50 μg/mL) with shaking at 37°C for at least 4 h
   c. Prepare 200 mL NZYM media in flask, take 1 mL as reference, and add 4 mL of culture into the flask, incubate with shaking at 37°C; monitor OD600 every hour until OD=0.6
   d. Collect 1 mL of sample as uninduced control; to the remaining culture, add IPTG (2 mM) to induce, let incubate with shaking at 37°C for 2 h
2. Protein extraction
   a. Spin down uninduced sample at 4000 rpm (Eppendorf 5414D) or 1500 *g* at room temperature (RT) (20°C–25°C) for 10 min; lyse directly in 50 μL of 2× sample buffer (diluted with ddH_2O from 4× Bolt LDS sample buffer, ThermoFisher B0007, containing reducing agent, Thermo B0009), boil for 5 min, store supernatant for later
   b. After 2 h of induction, collect bacteria pellet by spinning at 4000 rpm at 4°C for 10 min, discard supernatant; save 1 mL of induced sample, prepare supernatant as in preparation two steps 5 and 6 for uninduced sample
   c. Re-suspend with 10 mL TEN buffer (10 mM Tris-HCl pH7.5, 2 mM EDTA, and 150 mM NaCl) supplemented with fresh DTT (5 mM), lysozyme (0.1 mg/mL)
   d. Incubate on ice for 10 min
   e. Sonicate on ice with a microtip sonicator, 6 times at level 54, 30 s each cycle at 90% duty cycle
   f. The sample will still look cloudy, spin down at 12,000 rpm in SS-34 rotor at 4°C for 30 min, save the supernatant ("S1"). Make 500 μL aliquots, then either flash freeze in liquid N_2 or continue to preparation one, step 3 ProteinA-MNase purification
3. ProteinA-MNase purification
   a. Prepare IgG Sepharose 6 Fast Flow resin (Sigma GE17-0969-01): for each 500 μL of S1, aliquot 30 μL of bed volume of IgG resin into low binding tubes, add 1 mL of TEN buffer, spin down at 1,000 rpm for 1 min, repeat for 2 washes
   b. Either use S1 from preparation one, step 2 protein extraction, sub-step f, or thaw aliquot on ice, add to prepared IgG resin, incubate for 3 h at 4°C with rocking
   c. After incubation, save an aliquot of flow-through for preparation one, step 4 testing and quantification, wash resin twice with 1 mL of TEN supplemented with Empigen (0.03%, Sigma 30326), then twice with 1 mL of TEN-500 (TEN buffer containing 500 mM NaCl and 0.03% Empigen); each wash is 5 min incubation with rocking at 4°C followed by 1 min spin at 1,000 rpm
   d. After final wash, add 1 mL of NH_4Ac (5 mM, pH 5), directly followed by 1 min spin at 1,000 rpm
   e. Elute by adding 60 μL of HAc/NH_4Ac (0.5M, pH 3.4), incubate with rocking at 4°C for 10 min
   f. Spin at 1,000 rpm for 1 min, carefully take out supernatant and add into a tube containing 50 μL of 1% NaOH to balance the pH
   g. Add glycerol to 20% final concentration, aliquot and flash freeze to store at −80°C for up to 3 months.
4. Testing and quantification
   a. Run in an SDS-PAGE gel: uninduced and induced culture from preparation two, steps 5 and 6; S1; flow-through; one aliquot of eluate; leftover IgG resin; all boiled in 2× LDS sample buffer for 5 min and spin down at top speed, with supernatant saved
   b. Quantify by running BSA standards along with the samples

Note: If eluate is yellowish-green when 2× LDS sample buffer is added, it is too acidic and more 1% NaOH needs to be added

Note: Please also refer to Protocol step 5 – Quality Control.

Troubleshooting 1

Preparation two: Test digitonin concentration

5. Prepare 5% stock concentration of Digitonin in water.
6. Test cell of interest by re-suspending $1*10^5$ cells in Wash Buffer (20 mM HEPES pH7.5, 150 mM NaCl, 0.5 mM Spermidine) containing Digitonin, final concentration from 0.02–1%
7. Incubate at room temperature (20°C–25°C) for 1 min
8. Dilute cell suspension 1:1 with 0.4% Trypan blue
9. Check under the microscope the percentage of Trypan blue-positive cells. Use the lowest concentration of digitonin that resulted in over 75% Trypan blue-positive cells.

Preparation three: Analysis software environment setup

For this analysis, a custom Cut&Run Analysis Pipeline (CnRAP) was developed and uploaded to GitHub (See key resources table for URL). Refer to the GitHub repository for detailed installation and setup instructions which must be completed prior to analysis. Take note that for deeply sequenced libraries and/or for processing many samples, performing the computational processing on a computing cluster will drastically reduce the time required for processing.

⚠ CRITICAL: CnRAP consists of 6 scripts, numbered 01–06. Once the CnRAP environment is setup and genome indexes have been built for both intended genomes (human and yeast), a number of scripts require configuration for the user's specific system setup and environment. For detailed information regarding the required configuration of each script, refer to the GitHib Repository.

Note: The bioinformatics workflow described in this protocol has two assumptions. The first being that CnRAP has been setup as per the instruction outlined on the GitHub repository, while the second is that all scripts have been configured as outlined in the GitHub repository

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
SALL4 (2.5 ug used per sample)	Cell Signaling Technology	Cat# D16H12
Normal Rabbit IgG (2.5 ug used per sample)	Abcam	Cat# Ab171870
H3K9Me3 (D4W1U) (1 ug used per sample)	Cell Signaling Technology	Cat# 13969
Histone 3 (1 ug used per sample)	Abcam	Cat# Ab1791
Bacterial and virus strains		
BL21 (DE3)	Novagen	Cat# 69450
Chemicals		
FBS	Sigma	Cat# F2442
RPMI	Thermo Fisher	Cat# 11875119
DMEM	Thermo Fisher	Cat# 11965118
Trypsin-EDTA (0.25%)	Thermo Fisher	Cat# 25200114
Concanavalin A beads	Bangs Laboratories	Cat# BP531
proteinA-micrococcal nuclease	This paper	
Digitonin	Sigma	Cat# D141

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
NEBNext Ultra II DNA Library Prep Kit	NEB	Cat# M0541
NEBNext MμLtiplex Oligos for Illumina (index primers set 1)	NEB	Cat# E7335
Pippin Prep DNA Size Selection Kit	Sage Science	CDF3010
Experimental models: Organisms/Strains		
SNU-398	ATCC	Cat# CRL-2233
SNU-387	ATCC	Cat# CRL-2237
HeLa	ATCC	Cat# CCL-2
Software and algorithms		
enoLOGOS (no version number provided)	(Workman et al., 2005)	http://www.benoslab.pitt.edu/cgi-bin/enologos/enologos.cgi
Cut & Run Analysis Pipeline (CnRAP) (Github Release)	(Bassal, 2019; Kong et al., 2021)	https://github.com/mbassalbioinformatics/CnRAP
CUT&RUNTools (bitbucket release)	(Zhu et al., 2019)	https://bitbucket.org/qzhudfci/cutruntools/src/master/
Trimmomatic v0.36	(Bolger et al., 2014)	http://www.usadellab.org/cms/?page=trimmomatic
BWA v0.7.17-r1188	(Li and Durbin, 2009)	http://bio-bwa.sourceforge.net/
SAMtools v1.5	(Li et al., 2009)	http://samtools.sourceforge.net/
Stampy v1.032	(Lunter and Goodson, 2011)	https://www.rdm.ox.ac.uk/research/lunter-group/lunter-group/stampy
Picard v2.21.2	(Broad Institute, 2019)	https://broadinstitute.github.io/picard/
deepTools v2.5.7 – bamCoverage	(Ramirez et al., 2016)	https://deeptools.readthedocs.io/en/develop/
bedtools v2.25.0	(Quinlan and Hall, 2010)	https://bedtools.readthedocs.io/en/latest/
SEACR v1.1	(Meers et al., 2019)	https://github.com/FredHutch/SEACR
ChIPSeeker v1.20.0	(Yu et al., 2015)	https://guangchuangyu.github.io/software/ChIPseeker/
HOMER v4.10	(Heinz et al., 2010)	http://homer.ucsd.edu/homer/
ImageJ v1.51	(Schneider et al., 2012)	https://imagej.nih.gov/ij/
MEME v5.0.5	(Machanick and Bailey, 2011)	http://meme-suite.org/tools/meme-chip
R v3.6.1	(R Core Team, 2021)	https://www.R-project.org
Python2 (v2.7) and Python3 (v3.6.1)	(Python Software Foundation, 2021)	http://www.python.org
Perl v5.22	N/A	https://www.perl.org/
GenePix Pro v7.2	Molecular Devices	https://support.moleculardevices.com/s/article/GenePix-Pro-7-Microarray-Acquisition-Analysis-Software-Download-Page
Masliner	(Berger et al., 2006; Dudley et al., 2002)	http://arep.med.harvard.edu/masliner/supplement.htm
Universal PBM Analysis Suite	(Berger and Bulyk, 2009)	http://thebrain.bwh.harvard.edu/PBMAnalysisSuite/indexSep2017.html

STEP-BY-STEP METHOD DETAILS

Step 1: Cell nuclei purification

Timing: 1.5 h

This step allows for maximized binding of antibodies to nuclear factors and will result in cleaner CUT&RUN signal compared to protocol using whole cells

1. Cell collection and swelling
   a. Collect $2*10^6$–$2*10^7$ cells of interest by either scraping or centrifugation. We successfully used this protocol on SNU398, K562, and NB4 cell lines.

b. Resuspend in 5 mL of cold PBS and spin at 2,500 rpm at 4°C for 15 min
c. Measure the packed cell volume (PCV) of the cell pellet and add 5× PCV of 1× Buffer A (10× Buffer A contains 100 mM HEPES, 15 mM $MgCl_2$, 100 mM KCl, adjust pH to 7.9 with 10M KOH)
d. Vortex on high for 2s
e. Incubate on ice for 20 min to swell the cells

2. Nuclei collection
   a. Spin cell pellet at 2,500 rpm for 10 min
   b. Measure PCV again (should be slightly larger), add 2× PCV (original) of 1× Buffer A containing fresh protease inhibitor (Sigma complete protease inhibitor cocktail, 11697498001), 1 mM PMSF, 1 mM benzamidine, and 1 mM DTT; keep on ice
   c. Dounce with a 1 mL, prechilled glass douncer to break up the cell pellet, 7 strokes; alternatively, pass the cell lysate through a 25G needle 10×
   d. Incubate on ice for 10 min, transfer to 1.5 mL tube
   e. Spin down at 6,000 rpm at 4°C for 20 min
   f. Take out supernatant (cytoplasmic portion) for testing if needed; keep the pellet containing the nuclei
   g. Either flash freeze in liquid N_2 or move directly to the CUT&RUN experiments

Note: Each CUT&RUN reaction requires $2*10^6$ equivalent of cell nuclei.

Step 2: Antibody binding

Timing: 1.5 h to Overnight (16–18 h)

3. Resuspend nuclei in 1 mL room temperature (20°C–25°C) Wash Buffer (20 mM HEPES pH7.5, 150 mM NaCl, 0.5 mM Spermidine) supplemented with fresh protease inhibitor
4. Prepare Concanavalin A beads (Bangs Laboratories BP531)
   a. Take out enough beads for 10 μL/condition, can process together in one 1.5 mL low-binding microcentrifuge tube until protocol step 2.4.g
   b. Add 1.5 mL Binding Buffer (20 mM HEPES pH7.9, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$)
   c. Place in a magnetic stand to clear for between 30 s to 2 min
   d. Remove from the stand and add 1.5 mL Binding Buffer, mix by inversion
   e. 5s spin at 500 rpm
   f. Put the tube back to the stand to clear
   g. Re-suspend with 1:1 volume of Binding Buffer (i.e., 10 μL for 10 μL of beads) and aliquot into corresponding CUT&RUN low-binding tubes
5. While gently vortexing prepared Concanavalin A beads, add the nuclei in wash buffer from protocol step 2.3
6. Rotate at room temperature (20°C–25°C) for 10 min
7. Place on magnetic stand to clear (may take up to 1 min)
8. Take tubes off of the stand, add 50 μL of Antibody Buffer (Wash Buffer supplemented with fresh protease inhibitors, 2 mM EDTA, and 0.02%–0.1% Digitonin, exact concentration as determined in Preparation Two)
9. To each condition, add the appropriate amount of antibody or corresponding normal isotype IgG control (typically 2.5 μg of each, but exact concentration should be determined experimentally)
10. Incubate 1 h to overnight (16–18 h) with rocking at 4°C

⚠ CRITICAL: If you are running CUT&RUN for the first time, include a histone mark antibody for the quality control step.

Step 3: ProteinA-Mnase binding and Mnase digestion

Timing: 2 h

11. Quick spin the tubes at 500rom and place on magnetic stand to clear
12. Add 1 mL of Dig-Wash buffer (Wash Buffer with Digitonin, important: NO EDTA)
13. Mix by inversion, place on magnetic stand to clear
14. To each tube, add 50 μL of Dig-Wash containing 700 ng/mL of proteinA-MNase (concentration as determined in Preparation One, step 4)
15. Rotate for 1 h at 4°C
16. Quick spin the tubes at 500 rpm and place on magnetic stand to clear
17. Add 1 mL of Dig-Wash buffer
18. Repeat protocol steps 1, 3.16, and 3.17 for a total of 2 washes
19. Add 100 μL of Dig-Wash buffer along the sides to dislodge the beads
20. Place tubes in heat block sitting on wet ice (should be around 0°C)
21. While gently shaking each tube, add 2 μL of 100 mM $CaCl_2$
22. Incubate for 10–30 min (start with 30 min, if too long, then decrease the incubation, see Quality Control in step 5.47)
23. Add 100 μL of 2× Stop Buffer (0.34M NaCl, 20 mM EDTA, 4 mM EGTA, 0.02% Digitonin, 50 μg/mL RnaseA, 50 μg/mL glycogen, and 20 pg/mL heterologous spike-in purified DNA from yeast or bacteria)
24. Gently vortex
25. Incubate at 37°C for 10 min to release fragmented DNA
26. Centrifuge at 16,000 *g* for 5 min at 4°C
27. Place on magnetic stand to clear
28. Save clear supernatant containing DNA to a new tube

Step 4: DNA purification

Timing: 2.5 h to Overnight (16–18 h)

29. To each tube (~200 μL total), add 2 μL of 10% SDS (final concentration 0.1%), 2.5 μL of proteinase K (20 mg/mL)
30. Mix by inverting and incubate for 10 min at 70°C
31. Add 300 μL of phenol:chloroform:isopropanol (25:24:1) to each tube
32. Vortex for 2s
33. Transfer mixture to a phase-lock tube (Qiagen MaXtract, prepared by pre-spinning at 13,000 rpm for 30 s to settle the resin)
34. Centrifuge at 16,000 *g* for 5 min at room temperature (20°C–25°C)
35. Take out the top aqueous layer and transfer to a tube
36. Add 2 μL of glycogen (2 mg/mL, Thermo catalog number R0561)
37. Add 750 μL of cold 100% ethanol
38. Mix by inversion, incubate for at least an hour (up to overnight (16–18 h) at −20°C
39. Centrifuge at 16,000 *g* for 10 min at 4°C
40. Pour off liquid and dry on a piece of tissue
41. Rinse the pellet once with 950 μL of 100% ethanol
42. Centrifuge at 16,000 *g* for 1 min at 4°C
43. Pour off the liquid and drain on tissue
44. Air dry for 3 min
45. Dissolve the pellet in 25 μL of 1 mM Tris-HCl (pH8) and 0.1 mM EDTA (i.e., 0.1× TE)

Step 5: Quality control

Timing: 1 h

46. Quantify 1 μL of purified DNA by Qubit HS DNA kit (ThermoFisher Q32851) following manufacturer's instructions

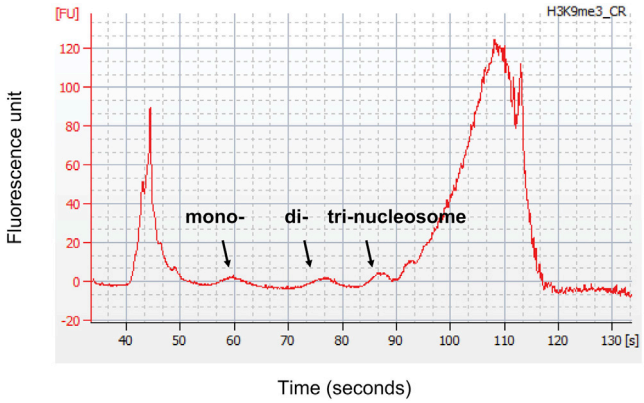


Figure 1. Expected Bioanalyzer electropherogram for CUT&RUN with an antibody against an histone mark

When running a histone mark CUT&RUN in parallel to your transcription factor of interest as quality control, one should expect to see small peaks representing mono-, di-, and tri-nucleosomes, which indicate successful proteinA-Mnase digestion.

Example amounts of DNA recovered
Histone mark: 15–20 ng/μL
TF: 10–15 ng/μL
IgG isotype control: <10 ng/μL

47. Run control histone mark CUT&RUN DNA on the 2100 Bioanalyzer system (Agilent)
48. Transcription factor CUT&RUN DNA are typically around 50–150 bp and may not show up on bioanalyzer, proceed to library amplification and size-selection

Note: Transcription factor CUT&RUN DNA may not show up on the bioanalyzer electropherogram if run prior to amplification. This is due to the low concentration of DNA present. To asses protocol efficacy, it is recommended to run, in parallel, a control CUT&RUN with an antibody against a histone mark.

⚠ CRITICAL: If running a control histone mark CUT&RUN, should see mono-, di-, and tri-nucleosomes in the Bioanalyzer traces (Figure 1).

Troubleshooting 2

Troubleshooting 3

Step 6: Library construction (with NEBNext ultra II DNA library Prep, NEB 7103, for transcription factor CUT&RUN DNA

Timing: 3 h

Note: For histone mark CUT&RUN library preparation, see Skene et al., 2018.

Note: This step is based on the protocol (https://dx.doi.org/10.17504/protocols.io.bagaibse) and publication (Liu et al., 2018)

49. Dilute CUT&RUN DNA (6 ng) into 25 μL with water, add 1.5 μL NEBNext Ultra II End Prep enzyme mix and 3.5 μL of NEBNext Ultra II End Prep reaction buffer (30 μL total)
50. Mix well and place in thermocycler, with heated lid set to >60°C: 30 min at 20°C, 60 min at 50°C, hold at 4°C
51. Dilute the adapters supplied by the NEBNext kit from the original 15 uM to 3 uM
52. Combine End repair mix from protocol step 6.50 with 15 μL of Ligation mater mix, 0.5 μL enhancer, 1.25 μL diluted adapter (~47 μL total)

53. Incubate for 15 min at 20°C in a thermocycler with no heated lid
54. Add 1.5 μL of USER enzyme to the mix
55. Mix well and incubate for 15 min at 37°C with heated lid set to >47°C
56. Vortex Ampure beads (Beckman A63880) and bring up to room temperature (20°C–25°C) for at least 30 min
57. Add 80 μL of beads (~1.75× volume) to the DNA mix from protocol step 6.55
58. Pipette 10 times to mix, incubate for 5 min at room temperature (20°C–25°C)
59. Place on magnetic stand for less than 5 min or until clear
60. Remove and discard the supernatant while the tube in the on the stand, leave the beads
61. While on the stand, wash twice with 200 μL of freshly made 80% ethanol, each time with incubation for 30 s at room temperature (20°C–25°C)
62. Remove all trance of ethanol with a p10 pipette tip
63. Air-dry the beads up to 5 min (don't over dry)
64. Remove the tubes from the stand and add 14 μL of 0.1× TE (same as protocol step 6.45)
65. Mix well or gently vortex
66. Incubate for 2 min at room temperature (20°C–25°C)
67. Place on the magnetic stand for 5 min
68. Remove and save 13 μL of supernatant containing DNA

Pause point: Can store at −20°C before PCR amplification

69. Set up PCR by combining 13 μL of DNA from protocol step 6.68, 15 μL NEBNext Ultra II Q5 Master Mix, 1 μL Index primer, and 1 μL Universal PCR primer (30 μL total)
70. Run PCR:
    Cycle 1: 98°C for 30 s
    Cycle 2: 57°C for 10 s, repeat 12 times total (Low Tm for transcription factor libraries)
    Cycle 3: 65°C for 5 min
    Hold at 4°C
71. Vortex Ampure beads and bring up to room temperature (20°C–25°C) for at least 30 min
72. Removing DNA products of >350 bp: add 24 μL (0.8×) Ampure beads to the PCR reaction and mix well, incubate for 5 min at room temperature (20°C–25°C), place the tubes on the magnetic stand for 5 min or until clear, carefully transfer the supernatant containing small DNA into a new tub3e
73. Keep DNA products of <150 bp: add 12 μL (1.2×) of re-suspended Ampure beads to the supernatant (so the effective bead to DNA ratio is 2×) from protocol step 6.72, mix 12 times, incubate for 5 min at room temperature (20°C–25°C), place the tubes on the magnetic stand for 5 min or until clear, remove and discard the supernatant
74. Wash twice with 200 μL of freshly made 80% ethanol, incubate for 30 s each time at room temperature (20°C–25°C)
75. Air-dry the beads for up to 5 min, do not over dry
76. Remove the beads from the magnetic stand and elute the DNA from the beads by adding 15 μL of 0.1× TE
77. Mix well and incubate for 2 min at room temperature (20°C–25°C)
78. Place the tubes back on the magnetic stand for 5 min or until clear
79. Transfer 13 μL of supernatant containing the library into a new PCR tube
80. Check size distribution on the 2100 Bioanalyzer system (Figure 2)

Pause point: Store the libraries at −20°C until sequencing

⚠ CRITICAL: Do not discard supernatant from protocol step 6.72 which contain the desired DNA library, discard the Ampure beads from this step which contain larger DNA fragments.

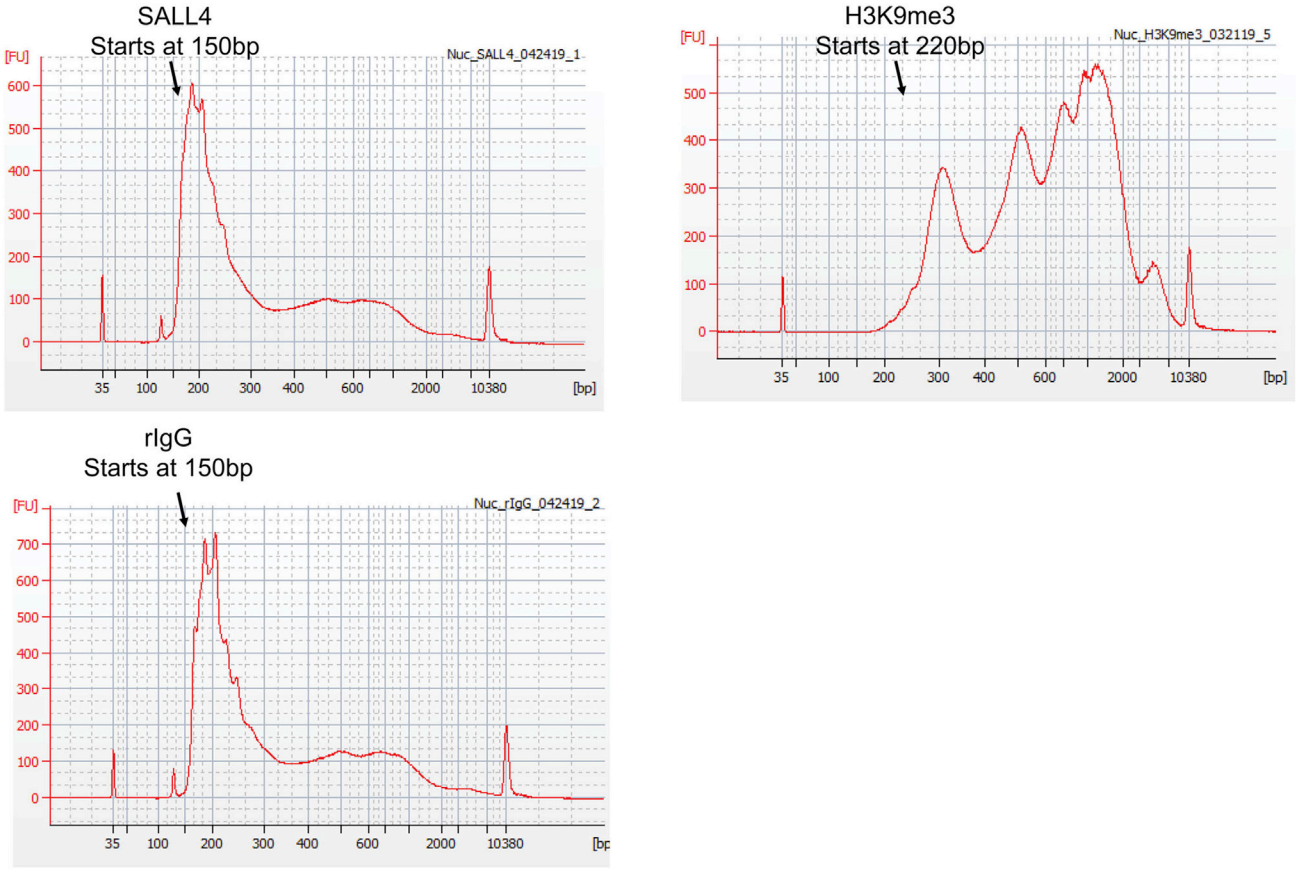


Figure 2. Expected Bioanalyzer electropherograms following library construction for nuclear DNA

Following library amplification, the bioanalyzer electropherograms should show sharp peaks at the size of the fragmented DNA with adapters as shown. Transcription factor CUT&RUN samples tend to show a smaller peak if following the described protocol

Step 7: Sequencing with MiniSeq (illumina): 4 h

81. Up to 24 barcoded quantified libraries can be mixed at equal molar ratio
82. (Optional) Remove PCR dimers with Pippin prep size selection according to the manufacturer's instructions
83. Perform paired-end sequencing (2 × 42 bp) on MiniSeq (Illumina) 5 million reads per library are sufficient

Step 8: Data analysis with CnRAP – time – hours to a couple of days depending on computer hardware, sequencing depth, and the number of samples to be processed

The bioinformatics pipeline written for this analysis consists of three essential scripts and three optional scripts. Each of the essential scripts performs a required function in processing the sequencing files in order to call the Cut&Run peaks while the optional scripts perform the peak annotation and motif enrichment analysis once the peaks are called. Briefly, the scripts perform the following functions:

- Script 01 – Read quality trimming and alignment to reference genomes
- Script 02 – Generation of bedgraph files and normalization
- Script 03 – Peak calling using SEACR
- Script 04 – (Optional) Annotate called peaks using ChIPSeeker
- Script 05 – (Optional) File preparation for HOMER motif enrichment analysis

- Script 06 – (Optional) File preparation for MEME motif enrichment analysis.

In this section, we will outline the function and purpose of each script, as well as outline how to call and use said script. Additional details can also be found on the GitHub page.

For those wishing to follow this analysis, the dataset analyzed in our original publication (Kong et al., 2021) is referenced in the Data Availability section. Users can download the raw and processed data from GEO and compare their results with the published findings.

> ***Note:*** Before beginning, it is important to note that for each sample sequenced, there should exist a pair of de-multiplexed fastq files corresponding to both read directions. These typically contain R1/R2 in the filename to denote read direction. Both files are needed per sample.

> ***Note:*** Ensure that you have followed preparation 3 in the "Before You Begin" section before proceeding. Next, the configured conda environment must be activated prior to attempting to run any of the following commands. See the GitHub repository for additional details.

84. The first step in running the analysis pipeline is to run CnRAP script 01 on each sample separately. Note, each sample will consist of 2 read files, R1 and R2, corresponding to the reads in both directions. This script takes care of performing all the necessary quality trimming steps required on the sequenced reads and alignment to the reference genomes. Firstly, trimmomatic is run to remove poor quality bases from both the start and the end of the reads while also removing any potential adapter sequences that might be found. Following this, the kseq_trimming tool is run to remove any additional barcode sequences. Once the reads are cleaned up, alignment to the reference genomes is performed using BWA followed by Stampy. In this protocol, reads are aligned to both the human (hg) and saccharomyces cerevisiae (sacCer) genomes. Alignment to hg is because the cell line used (SNU-398) is a human cell line, while alignment to the sacCer genome enables utilizing the spike-in DNA for normalizing. BWA and Stamy were chosen for alignment as they were found to result in improved alignment performance when benchmarked (Thankaswamy-Kosalai et al., 2017). Following alignment, unmapped reads are removed; bam files are sorted, indexed and alignment statistics are calculated.

To run script 01, use the following command.

```
> python3 01_cut_n_run_pairedReads_filter_align.py <sample_name> <read1_fq_gz> <read2_fq_gz> <num_cores> <aligned_folder>
```

Wherein:

> `<sample_name>` defines the prefix corresponding this sample;

> `<read1_fq_gz>` is the full path location of the read1 fastq file;

> `<read2_fq_gz>` is the full path location of the read2 fastq file;

> `<num_cores>` defines the number of processor cores to use for processing;

> `<aligned_folder>` is the folder where all output will be saved.

Following the above python3 call, a bash script will be generated `01_cut_n_run_pairedReads_filter_align.sh` which can simply be run in the terminal to perform the required steps for this stage of analysis.

Optional: Following genome alignment, some users may wish to assess the degree of PCR duplication present in their data. Marking and removal of PCR duplicates can be performed by Picard tools as well. Such an analysis can be performed using the following code. For additional information on interpreting the output of Picard, refer to the official documentation on the Broad website.

Optional: Following the generation of the aligned bam files, users can optionally run the tool "plotFingerprint" on the sample and IgG control bam files to check whether they see greater enrichment as expected in their sample. An example of such a figure can be found in Figure 3, wherein users can see a separation between the IgG and sample curves.

Optional: Following genome alignment, some users may wish to assess library complexity before proceeding. For this, Picard tools is an ideal and commonly accepted toolkit to do so. Such an analysis can be performed using the following code. For additional information on interpreting the output of Picard, refer to the official documentation on the Broad website.

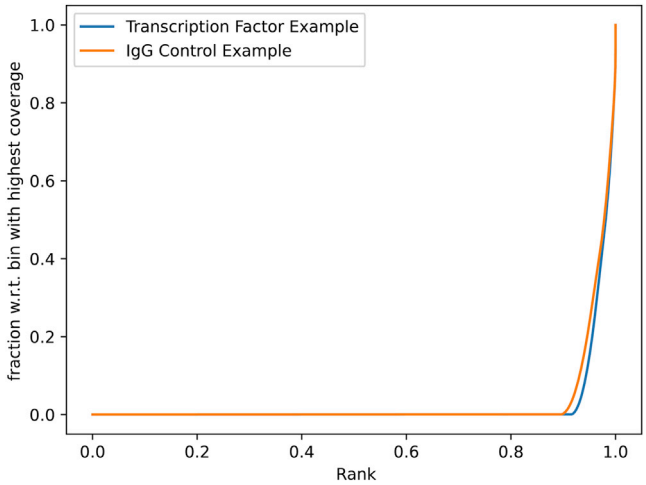


Figure 3. Example Enrichment fingerPrint plot following genome alignment

Following alignment to the human genome, users can optionally generate a global enrichment fingerPrint plot to assess the extent of enrichment observed in their sample over control. The sample curve (blue) should be closer to the bottom right corner than the control IgG curve (orange). The greater the separation between the curves the greater enrichment observed, which will likely result in more peaks being called at the peak calling step.

```
> java -jar picard.jar EstimateLibraryComplexity I=input.bam O=estimated_library_complexity_metrics.txt
```

```
> java -jar picard.jar MarkDuplicates I=input.bam O=marked_duplicates.bam M=marked_duplicates_metrics.txt
```

This command will simply mark the duplicate reads and not remove them. If you wish to remove duplicate reads, the `MarkDuplicates` command has two flags that can use utilized `REMOVE_DUPLICATES` and `REMOVE_SEQUENCING_DUPLICATES`. For full details on how to incorporate these flags into the afore mentioned command, refer to the official documentation on the Broad website.

Note: Be sure to set the same output `<aligned_folder>` for all samples as subsequent scripts will process all samples contained in given folders. Only script 01 is to be run separately per sample.

Troubleshooting 4

Troubleshooting 5

85. Next, CnRAP script 02 needs to be run on the output folder of script 01 which is supposed to contain the aligned bam files of all samples processed with script 01. Script 02 first takes the aligned bam files for each sample and converts them to bedgraph files. Next, the hg bedgraph files are normlized to the sacCer spike-in controls. For this, a normalization factor is calculated for each hg-sacCer file pair which is calculated as follows

$$normalization\_factor = \frac{10,000,000}{mapped\_reads\_per\_sacCer\_genome/2}$$

This normalization ensures that peak heights are adjusted correctly prior to peak calling in the next step.

To run script 02, use the following command.

```
> python3 02_cut_n_run_bamToBed_normalize_SEACRPrepv1.py <aligned_bams_folder> <normalized_beds_folder> <chrom_sizes_txt>
```

Wherein:

> <aligned_bams_folder> defines where the aligned bams, the output of script 01, are saved;

> <normalized_beds_folder> is the folder where the normalized bed files will be saved in preparation for running SEACR, the peak calling algorithm for Cut&Run;

> <chrom_sizes_txt> is a text file defining the base-pair size of each chromosome. This is required when converting the aligned bam files to bedGraph files. This file can be downloaded from UCSC directly.

Following the above python3 call, a bash script will be generated `02_cut_n_run_bamToBed_normalize_SEACRPrepv1.sh` which can simply be run in the terminal to perform the required steps for this stage of analysis.

86. Next, run CnRAP script 03 is required to be run over the normalized bedgraph files. This script proceeds to first make bigWig coverage files for all normalized bedgraph files (for viewing in IGV or on UCSC) and then calls peaks using SEACR in both "stringent" and "relaxed" modes.

To run script 03, use the following command

```
> python3 03_cut_n_run_SEACR_peakProcess_v1beds.py <seacr_location> <normalized_beds_folder> <output_folder> <chrom_sizes_txt>
```

Wherein:

> <seacr_location> defines where the Cut&Run peak caller SEACR is saved;

> <normalized_beds_folder> is the folder where the normalized bed files have been saved;

> <output_folder>is the folder where the called peak files will be saved;

> <chrom_sizes_txt> is a text file defining the base-pair size of each chromosome. This is required for bedGraph file manipulation.

Following the above python3 call, a bash script will be generated `03_cut_n_run_SEACR_peakProcess_v1beds.sh` which can simply be run in the terminal to perform the required steps for this stage of analysis.

87. Once peaks are called, users will likely want to annotate them although this is not essential. As such, script 04, is considered as an optional script to be run. If users wish to annotate their called peaks, script 04 can be utilized which will annotate called peaks using the R package, ChIPSeeker. In addition to annotating the called peaks, ChIPSeeker will also generate summary plots which may be of relevance to users. Script 04 will only annotate using ChIPSeeker. Alternatively, users may wish to annotate their called peaks using the HOMER annotatePeaks function. No script is provided as part of CnRAP to use HOMER though. For details on how to annotate your peaks using HOMER, refer to the official documentation.

    Optional: Once peaks are called, users may wish to perform motif enrichment analysis to investigate the presence of characteristic motifs. For this, users can use script 5 and/or script 6. Script 05 script, processes the called peaks and generates the required script to run motif enrichment analysis using the HOMER toolkit. Once generated, the corresponding bash script can be run in the terminal to perform said analysis. This script is optional as, depending on the experimental design, motif enrichment analysis may not be required. It is evident from user forums however that some users prefer to use the MEME-Suite of tools in preference to HOMER for motif discovery. As such, we have also provided script 06 which will prepare the required files for performing motif enrichment analysis using the MEME-Suite. The choice for which toolkit to use will be up to the users and their experiment. For the results generated using this protocol, both HOMER and MEME results were compared to ensure reproducibility in the called motif.

EXPECTED OUTCOMES

The first time an experimenter follows this protocol, they should run a parallel histone mark CUT&RUN library and should expect to observe mono-, di-, tri-nucleosome Bioanalyzer traces prior to library amplification that suggest (1) the purified proteinA-Mnase worked; and (2) the protocol worked. Once the bioinformatic pipeline was run, the experimenter should expect to see TF-specific peaks enriched above IgG-CUT&RUN levels. These peaks should be small enough to facilitate more precise identification of binding sites of transcription factors and their consensus binding sequences.

When assessing the sample quality using the bioanalyzer traces, if running in parallel a histone mark CUT&RUN, three sets of small peaks representing mono-, di-, and tri-nucleosomes can be expected, indicating that the proteinA-Mnase digestion was successful (Figure 1). Following library amplification and purification, the bioanalyzer traces expected should show a sharp peak at the size of the fragmented DNA with adapters (Figure 2).

Although Cut&Run does produce small DNA fragments, this does not necessarily pose an issue for the bioinformatics analysis. One point to be aware of though is that as part of our analysis pipeline, reads of length less than 20 bp are discarded. This is because shorter reads map ambiguously to the genome at a higher rate than longer reads and so are not deemed to be informative in this context.

For the bioinformatics analysis steps, a number of files can be expected as outputs of each script run. For CnRAP script 01, the key outputs for this script are aligned bam files and bam index files for each sample for both human and yeast genomes, i.e., 4 output files per input sample. For script 02, users should expect a single normalized bedGraph and a coverage bigwig file output per sample. The yeast genome data is used to normalize the human bedGraph file and so is no longer needed following normalization. Script 03 will generate two called peak files for each sample corresponding to SEACR's stringent and relaxed running modes, which output to use will depend on the biological

question being addressed. Script 04 will output an annotated peak file along with six summary plots showing the genomic binding distribution of the peaks per sample. Once annotated, you can load your alignment or coverage files in IGV or any Genome Browser and view your peaks. Example peaks can be found in Figure 4. Scripts 05 and 06 will each output a single folder per sample that contains all the files generated as part of the motif discovery analysis for that sample. Each folder will contain a hyper-text markup language (html) file to open in a browser to view the results.

LIMITATIONS

This protocol is dependent on the availability of a good and specific antibody against the transcription factor of interest, which can be tested empirically through western blotting and co-immunoprecipitation experiments. Furthermore, this protocol describes production of proteinA-Mnase, the activity of which should be tested by running a parallel CUT&RUN experiment with antibodies against histone marks. This parallel histone mark CUT&RUN experiment should be run as a positive control when performing CUT&RUN for the first time, see "Critical" under protocol step 2.

The pipeline developed for our analysis, CnRAP, is limited in its future use as the program Stampy, which is used in the sample alignment stage, is written in python2. As of 2021 python2 has reached its end-of-life and is no longer developed or maintained and Stampy is no longer maintained by its developers. While it is still possible to install python2 and Stampy, with instructions outlined on the CnRAP GitHub repository page, software incompatibility issues will likely arise in time. A future release of CnRAP will include enhanced features and will not have the same limitations as the release detailed in this protocol. While the exact future feature set is still developing, we are re-writing the analysis pipeline to be more robust and more user friendly in all aspects from setup, to usage and parameter setting. We also plan on developing a graphical interface to enable non-bioinformaticians to process their own Cut&Run datasets without needing to dive into the command-line, while still retaining the command-line for power users. When the new pipeline is developed, a link will be made available on the CnRAP GitHub repository.

TROUBLESHOOTING

Problem 1

Related to preparation 1, step 4

No purified proteinA-Mnase observed on gel.

Potential solution

If the induced but unpurified sample also does not contain the pAMnase, then this could be due to non-optimized induction, try longer induction period and different temperatures. If the protein is largely present in the flow through fraction, then incubate the bacterial lysates with 60 uL of IgG Fast Flow resin overnight (16–18 h) at 4°C with rocking.

Problem 2

Relating to protocol step 5: quality control

No distinct nucleosome peaks are seen in the parallel histone mark CUT&RUN sample.

Potential solution

This could be due to inefficient Mnase digestion, so one can use more Mnase digest and/or incubate for longer at 37°C.

Problem 3

Relating to protocol step 5: quality control

No TF CUT&RUN DNA was recovered from Qubit quantification.

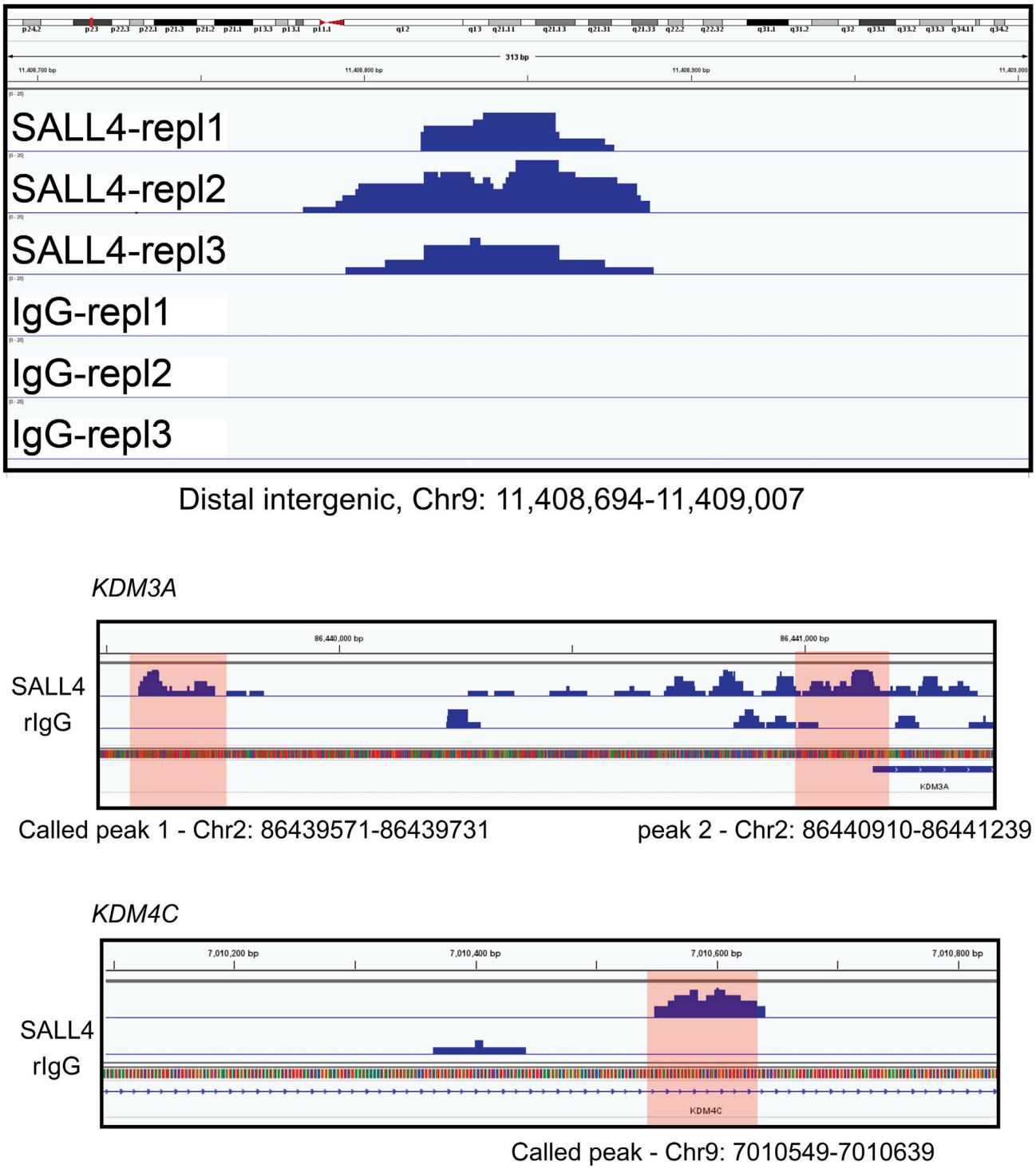


Figure 4. Example Cut&Run peaks at selected genomic loci

Once peaks are called, users can view the coverage files in UCSC genome browser or locally using a program such as IGV. At the loci where peaks are called, users should see noticeable enrichment in the sample compared to the IgG controls as seen here. The peaks shown here are from our published dataset reported in Cell Reports looking at SALL4 in the SNU398 liver cancer cell line.

Potential solution

The most likely answer is the antibody tested was not optimized for binding. Try immunoprecipitating with the TF-specific antibody and proteinA/G beads first by incubating it overnight (16–18 h) with nuclei that were purified following the protocol, lyse the nuclei the next day and perform western blotting to verify that the antibody can bind to the TF of interest to ensure the antibody is not the concern.

Problem 4

Relating to protocol step 8

Rarely, when running CnRAP script 01 (step 8.84), python2 will fail with a "broken pipe error".

Potential solution

This error typically occurred when assigning to python more system resources than are available. To resolve, simply re-run script 01 but assign fewer processing cores to the analysis.

Problem 5

Relating to protocol step 8

Pipeline scripts fail with the error "MemoryError".

Potential solution

This error would likely occur if your computer has run out of RAM and if it occurs will likely be when running script 01 (step 8.84). Having at least 16 Gb of RAM in your system available should be sufficient per sample though. This should be sufficient to run a single instance of script 01 (step 8.84) at a time. If you wish to run multiple samples simultaneously, or your samples are deeply sequenced, consider running the pipeline on a dedicated computing cluster.

If this error occurred you will need to provide more RAM for analysis. First try closing down any unnecessary programs running in the background to free as much RAM as possible. If this fails to work you will need to increase the size of your swap-disk. This is a much more complicated solution and you will need to Google how to do so for your own operating system and its version. The final solution is to simply find another computer with more hardware resources to perform the analysis on.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to Mahmoud A. Bassal (mahmoud.bassal@mymail.unisa.edu.au). Details and information regarding the computational analysis should also be directed to Mahmoud A. Bassal (mahmoud.bassal@mymail.unisa.edu.au).

Materials availability

All materials are available commercially or through AddGene with category numbers provided in-text.

Data and code availability

The datasets generated during our study are available on GEO with accession number GSE136332 and are referenced in the original study as well. Users can download the raw and processed data from GEO, run the outlined pipeline described herein, and compare the results. The bioinformatics pipeline utilized, CnRAP, is accessible on GitHub (https://github.com/mbassalbioinformatics/CnRAP).

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIH) grant number T32 HL066987 to N.R.K., grant number HL131477 to D.G.T. This work was further supported by the National Cancer Institute grant number R35 CA197697 to D.G.T.; the National Heart, Lung, and Blood Institute grant

number P01 HL095489 to L.C.; the Leukemia and Lymphoma Society grant number P-TRP-5855-15 to L.C.; and Xiu Research Fund to L.C. This work was also supported by the Singapore Ministry of Health's National Medical Research Council under its Singapore Translational Research (STaR) Investigator Award and by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence Initiative. We would also like to thank the Molecular Biology Core Facilities at the Dana Farber Cancer Institute for their help.

AUTHOR CONTRIBUTIONS

N.R.K. designed, performed, and optimized the SALL4 CUT&RUN. M.A.B. designed, wrote, and performed the bioinformatics analysis. N.R.K. and M.A.B. wrote and edited the protocol. L.C. and D.G.T. supervised the project and secured funding.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Bassal, M.A. (2019). Cut&Run Analysis Pipeline (CnRAP), ver. 1.0 (GitHub).

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat. Protoc. *4*, 393–411.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Broad Institute. (2019). Picard toolkit (Broad Institute, Broad Institute, Github repository).

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc. Natl. Acad. Sci. U S A *99*, 7554–7559.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Kong, N.R., Bassal, M.A., Tan, H.K., Kurland, J.V., Yong, K.J., Young, J.J., Yang, Y., Li, F., Lee, J.D., Liu, Y., et al. (2021). Zinc finger protein SALL4 functions through an AT-rich motif to regulate gene expression. Cell Rep. *34*, 108574.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, N., Hargreaves, V.V., Zhu, Q., Kurland, J.V., Hong, J., Kim, W., Sher, F., Macias-Trevino, C., Rogers, J.M., Kurita, R., et al. (2018). Direct Promoter Repression by BCL11A Controls the Fetal to Adult Hemoglobin Switch. Cell *173*, 430–442.e17.

Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936–939.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Meers, M.P., Tenenbaum, D., and Henikoff, S. (2019). Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. Epigenet. Chromatin *12*, 42.

Python Software Foundation (2021). Python Language Reference, ver. 2.7, 3.7, www.python.org.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

R Core Team (2021). R: A Language and Environment for Statistical Computing (Vienna, Austraia: R Foundation for Statistical Computing).

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675.

Skene, P.J., Henikoff, J.G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat. Protoc. *13*, 1006–1019.

Thankaswamy-Kosalai, S., Sen, P., and Nookaew, I. (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics *109*, 186–191.

Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D., and Benos, P.V. (2005). enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res. *33*, W389–W392.

Yu, G., Wang, L.-G., and He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382–2383.

Zhu, Q., Liu, N., Orkin, S.H., and Yuan, G.-C. (2019). CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis. Genome Biol. *20*, 192.