Genome-Wide Measurement and Computational Analysis of Transcription Factor Binding and Chromatin Accessibility in Lymphocytes

M. Firas Sadiyah[1,2] and Rahul Roychoudhuri[1,2]

[1]Laboratory of Lymphocyte Signalling and Development, Babraham Institute, Cambridge, United Kingdom
[2]Corresponding authors: *firas.sadiyah@babraham.ac.uk*; *rahul.roychoudhuri@babraham.ac.uk*

Cells of the adaptive immune system, including CD4$^+$ and CD8$^+$ T cells, as well as B cells, possess the ability to undergo dynamic changes in population size, differentiation state, and function to counteract diverse and temporally stochastic threats from the external environment. To achieve this, lymphocytes must be able to rapidly control their gene-expression programs in a cell-type-specific manner and in response to extrinsic signals. Such capacity is provided by transcription factors (TFs), which bind to the available repertoire of regulatory DNA elements in distinct lymphocyte subsets to program cell-type-specific gene expression. Here we provide a set of protocols that utilize massively parallel sequencing–based approaches to map genome-wide TF-binding sites and accessible chromatin, with consideration of the unique aspects and technical issues facing their application to lymphocytes. We show how to computationally validate and analyze aligned data to map differentially enriched/accessible sites, identify enriched DNA sequence motifs, and detect the position of nucleosomes adjacent to accessible DNA elements. These techniques, when applied to immune cells, can enhance our understanding of how gene-expression programs are controlled within lymphocytes to coordinate immune function in homeostasis and disease. © 2019 by John Wiley & Sons, Inc.

Keywords: activation • ATAC-seq • ChIP-seq • chromatin • differentiation • gene regulation • immune response • lymphocyte • T cell • transcription factor

How to cite this article:
Sadiyah, M.F., & Roychoudhuri, R. (2019). Genome-wide measurement and computational analysis of transcription factor binding and chromatin accessibility in lymphocytes. *Current Protocols in Immunology*, *126,* e84. doi: 10.1002/cpim.84

INTRODUCTION

Cells of the adaptive immune system possess the ability to undergo dynamic changes in population size, differentiation state, and function to counteract diverse and temporally stochastic threats from the external environment. Fundamental to this is the ability of lymphocytes to control gene expression in response to extrinsic cues and in a cell-type and differentiation-state-specific manner. Transcription factors (TFs) are proteins that bind to DNA and contribute to control of gene expression. TFs bind to DNA sequence motifs found within genomic regulatory elements such as gene promoters and enhancers. Although the promoter of an individual gene isoform is similar between distinct cell

CURRENT PROTOCOLS
A Wiley Brand

Current Protocols in Immunology e84, Volume 126
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpim.84

types, the repertoire of accessible enhancers controlling a given gene can be markedly different in distinct cell types, enabling distinct regulatory elements to control gene expression in various cell types. Thus, it is valuable to consider the spectrum of binding sites of a TF in the context of the available repertoire of accessible DNA elements within a given cell type.

Measurement of the distribution of proteins within the genome has been substantially aided by the development of ChIP-seq (Barski et al., 2007; Johnson, Mortazavi, Myers, & Wold, 2007; Mikkelsen et al., 2007; Robertson et al., 2007), which enables genome-wide TF binding sites to be determined. ChIP-seq allows measurement not only of TF binding but also of the distribution of histones bearing specific post-translational modifications. The presence of post-translational histone modifications affects chromatin accessibility and gene expression. Several methods have been developed to measure the distribution of accessible chromatin, including ATAC-seq (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013) and sequencing of DNase I hypersensitive sites (DNase-seq; Boyle et al., 2008, 2011; Hesselberth et al., 2009; Neph et al., 2012) or nucleosome-occupied regions such as micrococcal nuclease sequencing (MNase-seq; Ponts et al., 2010; Schones et al., 2008). ATAC-seq, in particular, provides a fast and sensitive means to profile accessible DNA loci and has low cell-number requirements. Tn5 introduces single-strand breaks (SSBs) in the DNA and integrates a template sequence for polymerase chain reaction (PCR)–based amplification in a process called tagmentation. During PCR amplification, primers can be used that add Illumina sequencing adapters and barcodes for multiplexed sequencing. DNA regions occupied by nucleosomes, bound by a single protein or a complex of proteins, or otherwise sterically inaccessible to Tn5, are protected from tagmentation and subsequent sequencing-based detection.

In this article, we provide step-by-step protocols to measure the genome-wide binding sites of TFs within lymphocytes (Fig. 1). We provide protocols for using ChIP-seq to map genome-wide TF binding sites and histone modifications (Basic

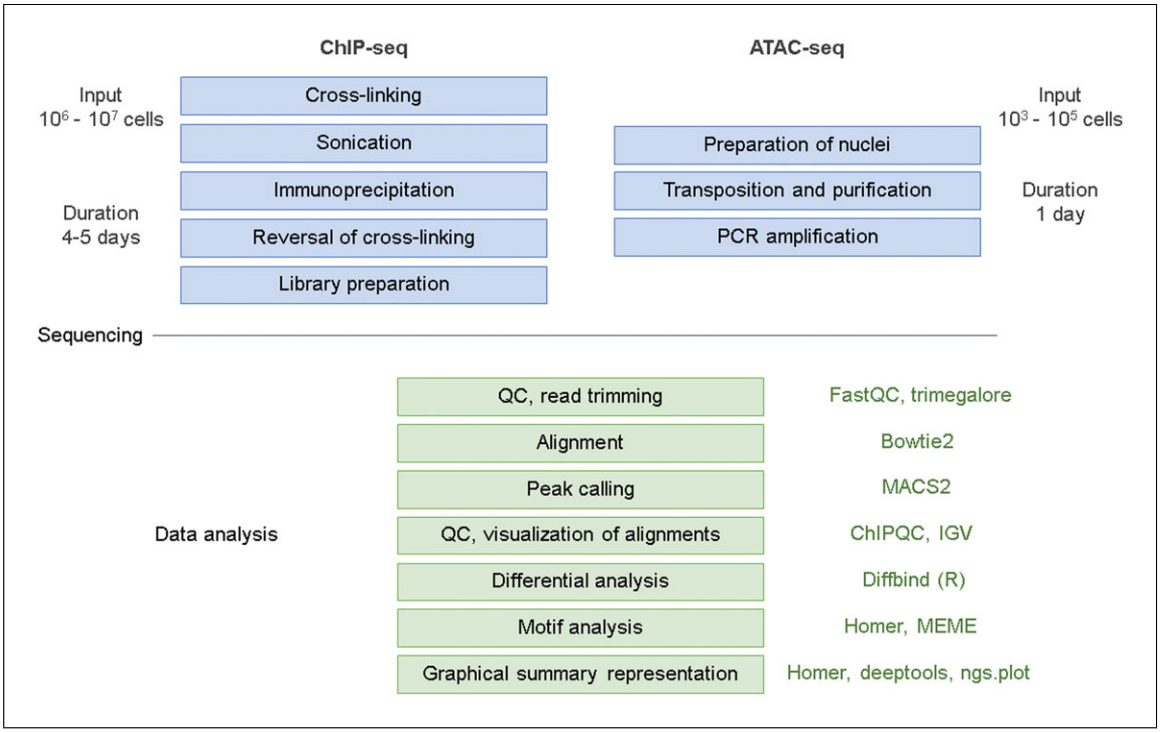

Figure 1 ChIP-seq and ATAC-seq analysis workflow.

Protocol 1) and to map the genome-wide repertoire of accessible DNA elements using ATAC-seq (Basic Protocol 2). We show how to perform quality control and align the sequencing data generated (Basic Protocol 3). Finally, we show how to analyze aligned data to map genome-wide TF binding sites and accessible DNA and determine differentially enriched/accessible sites (Basic Protocol 4). Critical to the success of these assays is the careful design of feasible strategies to obtain lymphocytes of sufficient purity and number to allow for an adequate number of experimental replicates, without substantial contamination from apoptotic or dead cells. We will consider the particular limitations and technical issues faced when studying lymphocytes and how to design experiments that are likely to produce informative and reproducible datasets.

CHROMATIN IMMUNOPRECIPITATION COUPLED WITH SEQUENCING (ChIP-seq) TO DETECT GENOME-WIDE TF BINDING SITES AND HISTONE MODIFICATIONS IN LYMPHOCYTES

BASIC PROTOCOL 1

This protocol describes a method to obtain DNA by immunoprecipitation using antibodies directed against TF proteins or histones. In ChIP, proteins are covalently cross-linked to DNA in order to maintain their location during processing. The cells are then lysed, and genomic DNA is sonicated to liberate shorter genomic fragments amenable to immunoprecipitation (IP) of protein−DNA complexes using antibodies. For the IP, a specific antibody conjugated to magnetic beads is then used to pull down protein−DNA complexes. The enrichment is followed by the reversal of the cross-linking to liberate DNA fragments, which are then purified for subsequent analysis. To measure binding of the immunoprecipitated protein to specific loci, and as an approach to validate the ChIP DNA prior to library preparation and sequencing, quantitative PCR–based measurement of ChIP enrichment at putative predicted or previously established binding sites is used, normalizing the enrichment signal to input signal and comparing enrichment with a negative control locus known not to bind the TF.

Choice of specificity controls in ChIP

The quality of the antibodies used to immunoprecipitate proteins of interest is fundamentally important to the validity of ChIP. Emphasis should be placed on selecting ChIP antibodies that have previously been validated in ChIP or ChIP-seq assays, or for which the specificity toward the protein of interest, preferably in native form, has been confirmed. To control for nonspecific enrichment and to determine the level of background signal, two types of controls are commonly used: (a) a nonspecific isotype control antibody, and (b) a 10% input control, which represents the fragmented chromatin that was available during the IP step. Signals from these control ChIP reactions can be used both in ChIP-PCR and ChIP-seq, as discussed further below.

If the opportunity is available, an alternative control to the isotype antibody control when immunoprecipitating TF proteins is use of a TF-specific antibody in wild-type cells and cells where expression of the TF protein has been disrupted. Loss of signal in TF-deficient cells provides robust evidence of the specificity of the assay, and this control should be considered for use in place of an antibody isotype control where possible, and where an antibody is being used for ChIP for the first time. Antibodies directed against tags fused to TF proteins of interest (e.g., anti-Flag antibodies directed against Flag-tagged TFs–Sigma anti-Flag antibody clone M2) can be used, in which case anti-Flag ChIP performed on wild-type cells should be used as specificity controls.

Generation of cells for ChIP assays

A high number of cells per replicate is required for ChIP assays (5−20 × 10^6 cells for TF ChIP and 2−20 × 10^6 cells for detection of histone post-translational modifications). In general, we recommend using four biological ChIP replicates per experimental condition if possible. It is a frequently encountered problem that the lymphocyte population of interest is rare, necessitating pooling of cells from multiple organisms for sufficient numbers of cells to be obtained. To ensure that cell populations of sufficient purity are obtained, we recommend using fluorescence-activated cell sorting (FACS)–based enrichment of lymphocyte populations, or magnetic selection, with post-sort confirmation of the purity (>95%) of the sorted cell populations.

Depending upon the frequency of the cell population of interest, it may be impractical to obtain sufficient numbers of relevant cells directly ex vivo. To surmount this, in vitro approaches to expand enriched lymphocyte populations of interest prior to the assay may be considered. In vitro culture also allows acute signals to be delivered to cells at specific time points prior to fixation and ChIP for the analysis of signal-dependent TF binding. Care should be taken to ensure that a limited frequency of dead cells is present in the culture prior to fixation and ChIP, as discussed in greater detail in Basic Protocol 2.

Materials

RPMI medium (Gibco, 11875093) supplemented with 10% heat-inactivated FBS, 1% Glutamax (Gibco, 35050061), and 1% penicillin-streptomycin (Gibco, 15140122)
Phosphate-buffered saline (PBS; Gibco, 10010023)
Cells of interest (B and T cell populations)
16% formaldehyde, methanol-free
Pierce^{TM} 16% Formaldehyde (w/v), Methanol-free; ThermoFisher, 28908
2.5 M glycine (Sigma, 50046) in PBS
Liquid nitrogen
Shearing buffer (see recipe)
20× EDTA/SDS (see recipe)
Dynabeads protein A (ThermoFisher, 10002D) *or* Dynabeads protein G (ThermoFisher, 10004D)
Antibodies: e.g.: anti-H3K4me1 (Abcam, ab8895) or anti-CTCF (Millipore, 07-729)
RIPA buffer (see recipe)
RIPA buffer plus 0.3 M NaCl (see recipe)
LiCl buffer (see recipe)
TE buffer, pH 8.0 (see recipe)
Triton X-100
10% (w/v) sodium dodecyl sulfate (SDS; Sigma, 71736)
20 mg/ml proteinase K (RNA grade; Qiagen, 19131)
Sodium chloride (NaCl)
DNA Clean & Concentrator-5 (Zymo, D4013)
NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, E7645S)
SPRIselect reagent (Beckman, B23317)
QIAquick PCR Purification Kit (Qiagen, 28104)
80% ethanol
NEBNext® Multiplex Oligos for Illumina® (Index Primers Set 1; NEB E7335S)
DNA LoBind Tubes (Eppendorf, 022431021)

50-ml conical centrifuge tubes
Refrigerated centrifuge and microcentrifuge
1.5-ml microcentrifuge tubes,
Branson 450-D sonicator
DynaMag-2 magnet (ThermoFisher, 12321D)
Tube rotator
65°C water bath

NOTE: Care should be taken not to introduce potential DNA contamination. Use deionized nuclease-free water where water is specified in all recipes and protocol steps.

ChIP and purification of ChIP-DNA

Prepare cross-linked cells

1. Warm RPMI medium to 37°C and cool PBS at 4°C.
2. Suspend 2 × 10^6 to 2 × 10^7 cells per ChIP replicate as appropriate for the target protein (e.g., use 2 × 10^6 for H3K4me1 ChIP and 2 × 10^7 for CTCF chip) in 10 ml pre-warmed RPMI medium in a 50-ml conical centrifuge tube.

   Care must be taken to use a consistent number of cells per replicate.

3. Add 665 µl of 16% formaldehyde to each sample for 1% final concentration. Use freshly opened methanol-free formaldehyde. Incubate at 37°C for exactly 10 min.

   Care should be taken not to extend the incubation time beyond 10 min, and that the incubation time is consistent across replicates.

4. Quench fixation reaction by adding 561 µl of 2.5 M glycine to a final concentration of 125 mM and incubate at room temperature for 5 min.
5. Add ice-cold PBS up to 50 ml, centrifuge 5 min at 1,105 × *g*, 4°C. Decant and vortex gently to loosen pellet.
6. Repeat the previous step.
7. Add 1 ml ice-cold PBS to the cell pellet. Transfer into a fresh 1.5-ml microcentrifuge tube.
8. Centrifuge 5 min at 1,105 × *g*, 4°C. Remove supernatant completely.
9. Snap-freeze cells in liquid nitrogen and store at –80°C.
10. This is a potential stopping point in the protocol.

Sonication to shear genomic DNA

11. Resuspend frozen pellet in 900 µl of shearing buffer.
12. Sonicate each sample using a Branson 450-D sonicator set at 36% amplitude with 15 cycles of sonication with each cycle lasting 20 s. Immerse tubes in ice during sonication—we have modified the lid of a 50-ml conical tube with a circular hole to accommodate the diameter of a 1.5-ml microcentrifuge tube such that the tube is held in place and immersed in ice that fills the 50-ml tube. Rest samples on ice for 30 s between sonication cycles. The sonication tip should be immersed two-thirds of the depth into the sample.

    Omission of SDS at this stage prevents excess formation of bubbles.

    We have optimized these conditions for T lymphocytes. If distinct cell types are to be used, we recommend performing a titration of sonication conditions on fixed samples and running clarified lysate on an agarose gel to assess size distribution.

13. Following sonication, add 47.4 µl of 20× EDTA/SDS to bring the concentrations of EDTA and SDS to match their final concentrations in RIPA buffer (see Reagents and Solutions) used for immunoprecipitation.
14. Vortex and incubate for 10 min on ice.

Perform immunoprecipitation

15. Microcentrifuge cell lysate 10 min at 13,000 rpm, 4°C. Carefully pipette 950 µl supernatant into a new tube, being careful not to disturb the pellet.

16. For each ChIP sample, separate 95 µl of the lysate into a new tube as ChIP input and store at 4°C.

17. Vortex Dynabeads before drawing up into a pipette using a precut pipette tip.

18. Add 40 µl of either Protein A or Protein G Dynabeads reactive against the ChIP antibody to a 1.5-ml tube.

19. Wash the beads twice, each time with 1 ml of PBS using a magnet stand to collect the beads during aspiration of wash supernatant.

    Avoid letting the beads dry at this stage

20. Remove supernatant using a magnetic stand. Add 100 µl of PBS with antibody, for example, use 6 µg of anti-H3K4me1 or anti-CTCF). Incubate tubes for 1 hr at room temperature with intermittent agitation.

21. Wash the antibody-bead conjugates by removing supernatant using a magnet stand. Wash twice, each time with 1 ml of PBS, while rotating for 5 min. Remove supernatant using a magnet stand.

22. Transfer cell lysates to the beads. Incubate at 4°C overnight on a rotator.

Washing of bead-bound protein−DNA complexes

23. Wash beads for 10 min while rotating at 4°C through the following wash cycles, removing supernatant after each wash using a magnetic stand to collect beads:

    Wash twice with 1 ml of RIPA buffer
    Wash twice with 1 ml of RIPA buffer + 0.3 M NaCl
    Wash twice with 1 ml of LiCl buffer
    Wash twice with 1 ml of TE + 0.2% Triton X-100
    Wash once with 1 ml of TE pH 8.0.

    Care must be taken not to allow beads to dry during removal of supernatant. To this end, we remove the supernatant and add the next wash buffer immediately to each tube individually. Solutions of wash media should be used at 4°C.

24. Resuspend beads in 100 µl of TE buffer, pH 8.0.

    Suspended beads can be stored indefinitely at −20°C, as well as the input samples from step 16; otherwise, continue to step 25.

Protein digestion of DNA-chromatin complexes

Please note that input samples need to be included in the following steps.

25. To each thawed sample from step 24, add 3 µl of 10% SDS. Add 5 µl of 20 mg/ml proteinase K, RNA grade. Incubate at 65°C for 4 hr with agitation.

26. Vortex briefly and spin down. Using a magnetic stand, transfer supernatant to a fresh tube.

27. Wash beads with 100 µl of TE buffer containing 0.5 M NaCl. Incubate at 65°C for 10 min with agitation.

28. Using a magnet stand, recover the supernatant with a pipette. Combine the supernatant from this step with that from step 26.

Purification of ChIP DNA using Zymo DNA Clean kit

29. Add five volumes of DNA binding buffer (from the Zymo DNA Clean & Concentrator kit) to 1 volume of sample. Mix briefly by vortexing. Transfer mixture to a provided Zymo-Spin column in a collection tube.

30. Centrifuge for 30 s at 10,000 to 16,000 × *g*. Discard the flow-through.

31. Add 200 μl DNA wash buffer (from the Zymo kit) to the column. Centrifuge at 10,000 × *g* for 30 s.

32. Repeat the previous step.

33. Put the column in a new 1.5-ml tube. Add 40 μl of TE buffer to the center of the column matrix. Incubate at room temperature for 1 min. Centrifuge for 30 s at 10,000 to 16,000 × *g*.

34. Before proceeding to library preparation (step 35), it is recommended that ChIP enrichment of positive and negative control regions in ChIP samples compared with input samples be determined by quantitative PCR (qPCR) to validate the sensitivity and specificity of the experiment.

Sequencing library preparation

The purpose of this section is to generate barcoded libraries for high-throughput sequencing from immunoprecipitated or input DNA generated in Basic Protocol 1 using Illumina-compatible adapters.

End repair, 5′ phosphorylation, and dA-tailing

35. Using PCR tubes, add 1× TE buffer to the DNA from step 34 to make up a total volume of 50 μl.

36. To each tube, add:

    a. 3 μl of NEBNext Ultra II end prep enzyme mix
    b. 7 μl of NEBNext Ultra II end prep reaction buffer
    c. Pipette up and down 10 times using a 100-μl repeat pipettor set at 30 μl

37. Perform thermal cycling on the reaction using the following conditions (with lid temperature set at >70°C):

    a. 30 min at 20°C.
    b. 30 min at 65°C.
    c. Hold at 4°C.

38. It is possible to stop here and store the reaction mixture at −20°C; however, ~20% loss may occur.

Performing adaptor ligation

39. To 60 μl of End Prep Enzyme Reaction mixture from the previous step, add the following (all items from NEBNext Ultra II kit):

    a. 30 μl NEBNext Ultra II Ligation Master Mix.
    b. 1 μl NEBNext Ligation Enhancer.
    c. 2.5 μl NEBNext Adaptor for Illumina,
    d. Pipette mix up and down 10 times using a 100-μl pipette.

40. Perform thermocycling on the reaction with lid heating off for 20 min at 15°C.

41. Add 3 μl of USER enzyme. Mix well.

42. Perform thermal cycling on the reaction with lid temperature set at >47°C for 15 min at 37°C.

43. It is possible to stop here and store the reaction mixture at −20°C.

Removal of free and self-ligated adaptors

44. Use the Qiagen QIAquick PCR Purification Kit to purify library DNA according to manufacturer's instructions, eluting in a final volume of 20 µl of water.

Size selection

The following size selection is for libraries with 200-bp inserts only. For libraries with different size fragment inserts, refer to Table no. 3.1 in the NEBNext Ultra II DNA Library Kit.

45. Warm SPRIselect beads at room temperature for 30 min. Mix the beads by vortexing.

46. Add 40 µl resuspended beads to the adaptor-ligated reaction (96.5 µl). Mix by pipetting up and down 10 times. Incubate 5 min at room temperature.

47. Place the tubes on the magnet for 5 min. Carefully transfer the supernatant—containing the desired DNA—into a new tube (do not discard supernatant). Discard the beads.

48. Add 20 µl resuspended beads to the supernatant and mix well. Incubate for 5 min at room temperature.

49. Place the tubes on the magnet for 5 min. Remove and discard the supernatant (contains unwanted DNA). Do not discard the beads (contains desired DNA).

50. Add 200 µl 80% ethanol while on the magnetic stand. Incubate for 30 s at room temperature. Remove and discard supernatant.

51. Repeat the wash step above.

52. Air dry the beads for 5 min while on the magnet stand with the lids open.

53. Add 17 µl of TE buffer, pH 8.0. Mix well by pipetting up and down ten times. Incubate for at least 2 min at room temperature. Place on the magnet for 5 min.

54. Transfer 15 µl to a new 1.5-ml microcentrifuge tube.

55. It is possible to stop here and store at −20°C.

PCR amplification of library DNA

56. To 15 µl of adaptor-ligated DNA fragments from the previous step 54, add:

    a. 25 µl NEBNext Ultra II Q5 Master Mix.
    b. 5 µl Index Primer/i7 Primer (from the NEBNext Multiplex Oligos for Illumina; use a distinct barcode for each sample to be run on the same sequencing lane and record the barcodes used for each sample).

       Alternatively, it is possible to use a combination of three different barcodes per sample to reduce the bias of sequencing efficiency for any one particular barcode (see Table 1).
    c. 5 µl Universal PCR Primer/i5 Primer (from the NEBNext Multiplex Oligos for Illumina).

57. Mix well by pipetting up and down 10 times.

Table 1 List of Sequencing Primers

Primer name	Primer sequence
Common primer	
Ad1_noMX	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
Barcode A	
Ad2.1_TAAGGCGA	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT
Ad2.2_CGTACTAG	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT
Ad2.3_AGGCAGAA	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
Barcode B	
Ad2.4_TCCTGAGC	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
Ad2.5_GGACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT
Ad2.6_TAGGCATG	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
Barcode C	
Ad2.7_CTCTCTAC	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT
Ad2.8_CAGAGAGG	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT
Ad2.9_GCTACGCT	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
Barcode D	
Ad2.10_CGAGGCTG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT
Ad2.11_AAGAGGCA	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT
Ad2.12_GTAGAGGA	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT
Barcode E	
Ad2.13_GTCGTGAT	CAAGCAGAAGACGGCATACGAGATATCACGACGTCTCGTGGGCTCGGAGATGT
Ad2.14_ACCACTGT	CAAGCAGAAGACGGCATACGAGATACAGTGGTGTCTCGTGGGCTCGGAGATGT
Ad2.15_TGGATCTG	CAAGCAGAAGACGGCATACGAGATCAGATCCAGTCTCGTGGGCTCGGAGATGT
Barcode F	
Ad2.16_CCGTTTGT	CAAGCAGAAGACGGCATACGAGATACAAACGGGTCTCGTGGGCTCGGAGATGT
Ad2.17_TGCTGGGT	CAAGCAGAAGACGGCATACGAGATACCCAGCAGTCTCGTGGGCTCGGAGATGT
Ad2.18_GAGGGGTT	CAAGCAGAAGACGGCATACGAGATAACCCCTCGTCTCGTGGGCTCGGAGATGT
Barcode G	
Ad2.19_AGGTTGGG	CAAGCAGAAGACGGCATACGAGATCCCAACCTGTCTCGTGGGCTCGGAGATGT
Ad2.20_GTGTGGTG	CAAGCAGAAGACGGCATACGAGATCACCACACGTCTCGTGGGCTCGGAGATGT
Ad2.21_TGGGTTTC	CAAGCAGAAGACGGCATACGAGATGAAACCCAGTCTCGTGGGCTCGGAGATGT
Barcode H	
Ad2.22_TGGTCACA	CAAGCAGAAGACGGCATACGAGATTGTGACCAGTCTCGTGGGCTCGGAGATGT
Ad2.23_TTGACCCT	CAAGCAGAAGACGGCATACGAGATAGGGTCAAGTCTCGTGGGCTCGGAGATGT
Ad2.24_CCACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTGGGTCTCGTGGGCTCGGAGATGT

58. Perform thermal cycling on the reaction using the following conditions:

1 cycle:	30 s	98°C	(initial denaturation)
15 cycles:	10 s	98°C	(denaturation)
	75 s	65°C	(annealing/extension)
1 cycle:	5 min	65°C	(final extension)
1 cycle:	indefinitely	4°C	(hold).

Cleaning PCR product and size selection

59. Use the Qiagen QIAquick PCR Purification Kit according to manufacturer's instructions to purify the amplified libraries.

60. Vortex SPRIselect beads.

61. Add 87 μl (0.9×) resuspended beads to the adaptor-ligated reaction (93.5 μl). Mix by pipetting up and down for 10 times. Incubate for 5 min at room temperature.

62. Place the tubes on the magnet for 5 min. Remove and discard the supernatant. Do not discard the beads.

63. Add 200 μl of 80% ethanol while on the magnetic stand. Incubate for 30 s at room temperature. Remove and discard supernatant.

64. Repeat the wash steps above.

65. Air dry the beads for 5 min while on the magnet stand with the lids open.

66. Add 33 μl of 1× TE buffer, pH 8.0. Mix well by pipetting up and down for ten times. Incubate at least 2 min at room temperature.

67. Place on the magnet stand for 5 min. Transfer 30 μl to a new 1.5-ml tube.

68. It is possible to stop here and store at −20°C.

BASIC PROTOCOL 2

ASSAY FOR TRANSPOSASE-ACCESSIBLE CHROMATIN USING SEQUENCING (ATAC-seq) ANALYSIS TO DETECT ACCESSIBLE CHROMATIN IN LYMPHOCYTES

In ATAC-seq, cell membranes are lysed and nuclei are isolated. Tn5 transposase is then added, which introduces SSBs and transposes templates for PCR amplification. DNA fragments are purified and amplified by PCR, incorporating indexing barcodes and Illumina adapter sequences so that samples can be multiplexed for sequencing. For ATAC-seq analysis, a significantly lower number of cells is required than when using ChIP-seq analysis (5×10^2 to 5×10^4 cells), permitting analysis of rare populations ex vivo. In general, we recommend using four biological replicates per experimental condition. The presented protocol must be performed on live cells; therefore, careful consideration should be exercised with respect to the timing of cell isolation so that sufficient time is allowed to proceed to at least step 9 of the ATAC-seq protocol below, at which point samples can be frozen and stored for further processing. In most cases, to ensure that cell populations of sufficient purity are obtained, we recommend using fluorescence-activated cell sorting (FACS)–based enrichment of lymphocyte populations. In vitro cultures may be used to allow for stimuli to be delivered to cells at specific timepoints prior to ATAC-seq analysis. In this case, since cells cannot be fixed or frozen prior to analysis, if stimulation time courses are to be performed, we recommend a "staggered-start" time course design if possible, allowing cells to be harvested simultaneously and placed on ice prior to initiation of the ATAC-seq assay.

Elimination of apoptotic cells prior to ATAC-seq analysis

Cell death is a homoeostatic process required for appropriate control of lymphocyte cell responses. As a result, lymphocytes undergoing stimulation in vivo or in vitro can undergo a high rate of physiological cell death. Accumulation of dead cells is particularly an issue with cultured lymphocytes, whose nonadherent nature limits the ability to easily remove apoptotic cells from the culture. Degradation of nuclear DNA into nucleosomal units is one of the hallmarks of apoptotic cell death, and such DNA, if present as a contaminant in ATAC-seq assays, forms a substrate for tagmentation by the Tn5 transposase and subsequent amplification. Such amplification contributes to background signal in ATAC-seq libraries, resulting in issues with library size normalization. Therefore, we recommend FACS-sorting cultured cells prior to ATAC-seq where apoptotic cells have accumulated to a frequency of >5% in culture. Since it is important that the number of cells used in ATAC-seq assays be similar across replicates and conditions, the precise number of cells required can be sorted directly into 1.5-ml microcentrifuge cells prior to ATAC-seq.

Materials

FACS-sorted fresh cells: 5×10^2 to 5×10^4 suspension cells per replicate
Phosphate-buffered saline (PBS; Gibco, 10010023)
ATAC-seq lysis buffer (see recipe)
Nextera DNA Library Preparation Kit (Illumina FC-121-1030)
MinElute PCR Purification Kit (Qiagen, 28004)
Index primers for ATAC-seq (see Buenrostro et al., 2013)
NEBNext high-fidelity 2× PCR Master Mix (NEB, M0541S)
QIAquick PCR Purification Kit (Qiagen, 28104)

1.5-ml microcentrifuge tubes
Refrigerated microcentrifuge
PCR tubes

Preparation of nuclei

1. Wash 5×10^2 to 5×10^4 FACS-sorted fresh cells in 1.5-ml microcentrifuge tubes with cold PBS.

   Care should be taken to use a consistent cell number across replicates.

2. Centrifuge for 5 min at 4°C and use a pipette to remove supernatant, ensuring minimal residual supernatant.

3. Resuspend cells in 100 μl of cold 1× ATAC-seq lysis buffer. Gently resuspend by pipetting up and down five times. Incubate the cells on ice for 10 min to allow the lysis of the cell membranes.

4. Centrifuge nuclei immediately 10 min at 500 × *g*, 4°C. Discard supernatant (cytoplasm) and keep pellet (nuclei) on ice.

Transposition reaction

5. To make the transposition reaction master mix, combine 25 μl of 2× TD buffer (Tagment DNA buffer) and 2.5 μl of TDE1 (Tagment DNA Enzyme) to 22.5 μl of RNase/DNase-free water per sample, allowing for reagent excess.

   Both of the reagents mentioned above are from the Nextera DNA Library Preparation Kit.

6. Distribute 50 μl of transposition reaction master mix into each tube containing a nuclear pellet (step 4) and resuspend by pipetting up and down five times.

7. Incubate the transposition reaction at 37°C for 30 min.

Purification of transposed DNA fragments

8. Use the Qiagen minElute PCR kit to purify transposed DNA and elute into 10 µl of elution buffer, according to manufacturer's instructions.

9. It is possible to stop here and store eluted DNA in elution buffer at −20°C for processing at a subsequent time.

Amplification of DNA fragments

10. To prepare PCR reactions, combine the following into PCR tubes:

    a. 5 µl of transposed DNA.
    b. 1 µl of 5′ and 3′ primer mix (a different index primer should be used for each sample to be run on the same lane).
    c. 27.5 µl of NEBNext High-fidelity 2× PCR Master Mix.
    d. 20.5 µl of RNase/DNase-free water (from NEBNext kit).

11. Perform thermocycling. In order to reduce PCR bias caused by overamplification of ATAC-seq libraries, determination of the optimal cycle number should be carried out as follows. Amplify samples for five cycles and take an aliquot of the PCR reaction as a template for further quantitative polymerase chain reaction (qPCR) using Sybr Green for a further 20 cycles. The number of cycles required, including the five initial cycles, to amplify the library during the linear amplification phase should be selected. Typically, 10 cycles are sufficient to amplify ATAC-seq libraries from 5×10^4 lymphocytes.

    Thermocycling conditions:

1 cycle:	5 min	72°C
1 cycle:	30 s	98°C
15 cycles:	10 s	98°C
	30 s	63°C
1 cycle:	1 min	72°C

Purification of PCR fragments

12. Use the Qiagen QIAquick PCR Purification Kit according to manufacturer's instructions to purify library DNA for sequencing. It is important in the elution step to ensure that the elution buffer is pipetted directly onto the center of the QIAquick membrane for complete elution of the bound DNA. Elute in 20 µl of water and store at −20°C. Proceed to paired-end sequencing (recommended; see Removing Duplicates below).

BASIC PROTOCOL 3

BIOINFORMATICS VALIDATION, NORMALIZATION, AND ALIGNMENT OF SEQUENCING DATA FROM ChIP-seq AND ATAC-seq EXPERIMENTS

Materials

Software

FASTQC 0.11.8 (*https://www.bioinformatics.babraham.ac.uk/projects/fastqc/*)
Trim GAlore 0.5.0 (*https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/*)
Bowtie2 2.3.2 (Langmead & Salzberg, 2012; *http://bowtie-bio.sourceforge.net/bowtie2/index.shtml*)
Samtools 1.9 (Li et al., 2009; *http://www.htslib.org/*)
Sambamba 0.6.8 (Tarasov, Vilella, Cuppen, Nijman, & Prins, 2015; *http://lomereiter.github.io/sambamba/*)
Bedtools 2.25.0 (Quinlan & Hall, 2010; *https://bedtools.readthedocs.io/en/latest/*)
atacqv 1.0.0 (*https://parkerlab.github.io/ataqv/demo/*)

Phantompeakqualtools 1.1 (Landt et al., 2012; *https://www.encodeproject.org/software/phantompeakqualtools/*)

Data

Raw Fastq sequencing files generated by Basic Protocols 1 and 2

Data processing

After sequencing the libraries generated in Basic Protocols 1 and 2, access the raw sequencing data generated in fastq format. Each fastq file represents a lane of sequencing. In the case of multiplexing, where more than one sample is loaded on the same lane, each fastq file represents one de-multiplexed sample. FastQ files store biological sequences along with their corresponding quality scores. The size of these files varies depending on read length and the sequencing equipment. In the case of paired-end sequencing, each sample is associated with two fastq files, usually denoted as Read 1 (R1) and Read2 (R2).

Initial quality assessment

Prior to the processing of FASTQ files, it is recommended to assess read quality. One of the most commonly used tools to perform quality checks is FASTQC. Among other checks, it provides insights into Phred quality scores across base pairs of sequenced reads, GC content, duplicate sequences, and over-represented sequences. Phred quality score represents the confidence in assigning each base called by the sequencer. In the context of next-generation sequencing, Phred quality scores range from 2 to 40, where a high score represents a high probability that the base called by the sequencing machine is correct. A score of 20 or above is considered acceptable, but it is recommended to keep reads of score 30 or above for further analysis.

```
fastqc seqfile1 seqfile2 .. seqfileN
```

Trimming adaptor sequences

During library preparation, forward and reverse primers incorporating indices (or barcodes) and Illumina adapter sequences are used to generate DNA fragments. Sequence information downstream of the 5′ adapter is attained. However, if the DNA fragment is shorter than the read length, then reads can run into the 3′ adapter, causing low mappability and misalignment to the underlying genome. Therefore, it is important to remove adapter sequences from short reads. Trim Galore is a wrapper that simultaneously runs Cutadapt and FastQC. It scans each read for the first 13 bp of Illumina standard adapters `AGATCGGAAGAGC` and cuts the matched sequence as well as all trailing sequences. It supports paired-end data when using the option `--paired` to keep forward and reverse reads files associated. To run Trim Galore, issue the following command:

```
trim_galore --paired \
            --fastqc_args “-q” seqfile1 seqfile2 .. seqfileN
```

By default `trim_galore` uses `--gzip` to compress the output and `--phred33` to instruct Cutadapt to use ASCII+33 quality scores as Phred scores (Sanger/Illumina 1.9+ encoding) for quality trimming. Additional FastQC arguments can be passed using the `--fastqc_args` option.

Aligning reads to a reference genome

After checking for the quality of the reads and scanning them for adaptor sequence, an aligner software such as Bowtie2 or BWA is used to search the reference genome for the regions that best match the read sequence. The output of Bowtie2 is in a text format known as sequence alignment/map format (SAM), which is uncompressed. Samtools can be used to produce a binary alignment map (BAM), which is compressed and requires less space for storage. Bowtie2 supports multiple threads that can be adjusted using the --threads option. To run Bowtie2, use the following command:

```
bowtie2 --threads 8 -X 2000
          -x genome \
          -1 read1.fastq -2 read2.fastq 2> out.bt2.log \
          | samtools view -b -q 30 - > output.bam
```

where

-x is the path to the genome index generated previously.
-1 is the path for the forward read fastq file.
-2 is used for the reverse read fastq file in a paired-end experiment.

Bowtie uses indexing as a computational strategy to speed up its mapping algorithm. It converts the reference genome into a Burrows-Wheeler transformed genome, which facilitates fast lookup of the query reads. Genome sequences in fasta format can be downloaded from UCSC (*http://hgdownload.cse.ucsc.edu/downloads.html*) or ENSEMBL (*ftp://ftp.ensembl.org/pub/release-95/fasta/*). For example, to download the mm9 assembly of the mouse genome from UCSC:

```
# download `chromFa.tar.gz ` from UCSC golden path
wget http://hgdownload.cse.ucsc.edu/goldenPath/mm9/bigZips/
chromFa.tar.gz

# uncompress the downloaded file
tar -xvzf chromFa.tar.gz

# remove `*random.fa` chromosomes
rm -rf *_random.fa

# concatenate all FASTA files into a single file
cat *.fa > mm9.fa
# index the concatenated .fa file using `samtools`
samtools faidx mm9.fa
```

To generate a bowtie2 index, run the following command:

```
bowtie2-build genome.fa genome_name
```

Links to frequently used index genomes are available on the Illumina iGenome website (*https://support.illumina.com/sequencing/sequencing_software/igenome.html*). For example, to analyze human and mouse sequencing samples, we use hg38 and mm9 reference assemblies, respectively.

It is important to bear in mind the difference in chromosome annotation between UCSC and ENSEMBL, where the former uses the suffix chr before the chromosome number and chrM to designate the mitochondria chromosome, while the latter uses numbers or MT to designate the chromosomes and the mitochondrial chromosome, respectively. Different downstream bioinformatics tools assume one annotation or the other, and manual conversion may be required between these two formats.

To index alignment BAM files, we can use samtools, which will generate an index file matching the original file name and ending with .bai file extension

```
Samtools index file.bam
```

Removing duplicates

Duplicates can be either natural or artificial. Natural duplicates are reads that result from identical fragments found in the original sample. Artificial reads, on the other hand, are reads that were produced artificially during library preparation, for example, during the PCR amplification step or library sequencing, i.e., optical duplication or cluster duplication. In specific applications, de-duplication is recommended to mitigate this bias. However, there is the risk of losing reads that represent natural duplicates. Many tools are available to de-duplicate samples including Picard and sambamba. To de-duplicate using sambamba, use the following command:

```
sambamba markdup \
        --remove-duplicates \
        --nthreads 8 \
        --hash-table-size 600000 \
        --overflow-list-size 600000 \
        input.bam output.dedup.bam
```

where

`--remove-duplicates` is used to remove the marked duplicates.
`--nthreads` is the number of threads to be used.
`--hash-table-size` is the size of the hash table for finding read pairs and for good performance; the value should be more than the average coverage multiplied by the insert size (default is 262,144 reads).
`--overflow-list-size` is the size of the overflow list where reads, thrown away from the hash table, get a second chance to meet their pairs; when the size is increased, fewer temporary files are created during the process (default is 200,000 reads).

Downsampling of library size

Differences in the depth of sequencing are a potential source of artifacts between samples. Optionally, one way to avoid this is to randomly downsample the size (in read counts) of all samples to the sample with the lowest library size. To calculate the lowest depth of sequencing across all samples, run the following command in the directory where bam files reside:

```
for file in *.bam; do samtools idxstats $file | cut -f3 \
| awk ‘BEGIN {total=0} {total += $1} END {print total}’;
  done \
| sort -n \
| awk ‘NR == 1’
```

In the above command, `samtools idxstats` is used to calculate the number of mapped reads per sample. Awk is then used to extract the total mapped reads per bam file (sample), and the lowest value across all bam files (samples) is reported. Next, the fraction by which the samples will be downsampled is calculated:

```
FRACTION = Total mapped reads (sample) / Lowest depth of
sequencing
```

Finally, Samamba is used to downsample the samples by passing the FRACTION calculated above to the `--subsample=` option. It is recommended to use a `--subsampling-seed` option and set the `SEED` for subsampling to ensure reproducibility.

```
sambamba view --subsample=FRACTION --subsampling-seed=SEED
--nthreads 2 -f bam input.bam -o downsampled.bam
```

ATAC-seq quality control

A method of quality control suited to ATAC-seq data is ATAC-seq QC and visualization (ataqv). It examines the aligned reads and provides metrics on mapping quality, the ratio of short to mononucleosome fragment counts, and reads mapping to autosomal and mitochondrial references.

To run atacqv on bam files from ATAC-seq, use the following command:

```
ataqv GENOME sample.bam --metrics-file sample.bam.ataqv.
json
```

For accurate duplication metrics, duplicate reads need to be marked in the bam files. In addition, BED files from peak callers such as MACS can be used to produce read coverage of the peaks. The result is a JSON file for each bam file.

Then, mkarv script is run to collect ataqv results and sets up a web application to visualize them which requires Python 2.7 or newer.

```
mkarv output_report sample.bam.ataqv.json
```

As a result of the above steps, the raw sequencing data has been assessed for sequencing quality and trimmed of adapter sequences. Reads have been aligned to the corresponding reference genome, removing reads with poor quality, duplicates have been optionally removed, and samples have been optionally downsampled in library size. The processed aligned BAM file can be used for subsequent analysis in Basic Protocol 4.

BASIC PROTOCOL 4

MAPPING OF GENOME-WIDE TF BINDING AND ACCESSIBLE CHROMATIN SITES, DNA SEQUENCE MOTIF ANALYSIS, AND DIFFERENTIAL PEAK ANALYSIS OF ChIP-seq AND ATAC-seq DATA

Materials

Software

MACS2 2.1.2 (Zhang et al., 2008) (*https://github.com/taoliu/MACS*)
Diffbind 2.10 (Ross-Innes et al., 2012) (*https://bioconductor.org/* packages/release/bioc/html/DiffBind.html)
NucleoATAC 0.3.4 (Schep et al., 2015) (*https://github.com/GreenleafLab/NucleoATAC*)

Data

Processed alignment BAM files from Basic Protocol 3

Genome-wide mapping of TF binding and accessible chromatin sites

Following de-duplication and downsampling of aligned reads, a peak caller is used to identify, within the samples, loci where reads are significantly enriched across the genome. A widely used peak caller is Model-based Analysis of ChIP-seq (MACS). MACS can operate on independent libraries or subtract signals from input control libraries (recommended) or signals from alternative experimental groups. MACS2 has two modes: a regular one that is recommended for narrow peaks such as TF binding sites or a number of histone marks such as H3K4me1 ChIP-seq, and a broad one that can be applied to more broadly distributed histone marks.

To apply MACS2 for narrow peaks, run the following command:

```
macs2 callpeak \
          --treatment sample.bam \
          --control control.bam \
          --format BAMPE \
          --gsize mm \
          --name sample_name
```

where

`-f` is the file format, for example, SAM, BAM, BAMPE (for paired-end BAM files).
`--gsize` is the size of the reference genome.
`--qvalue` is the minimum FDR cutoff to call significant regions (default is 0.05).
`--name` is the name of the sample.

To call MACS for broad peaks, run the following command:

```
macs2 callpeak \
          --treatment sample.bam \
          --control input.bam \
          --format BAMPE \
          --gsize mm \
          --broad \
          --broad-cutoff 1e-5 \
          --qvalue 0.01 \
          --name sample_name
```

where

`--broad` instructs macs to put nearby highly enriched regions, determined by `--qvalue`, into a broad region with loose cutoff, determined by `--broad-cutoff`.

To determine an appropriate cutoff, use `macs2 callpeak --cutoff-analysis`, without passing `--broad`. The number of peaks at distinct cutoff values will be reported enabling a decision regarding cutoff to be made.

For calling peaks from ATAC-seq data, MACS2 is instructed to build a shifting model and to pileup reads on the whole fragment by using `--format BAMPE`.

```
macs2 callpeak \
       --treatment sample.bam \
       --format BAMPE \
       --keep-dup all \
       --gsize mm \
       --broad \
       --broad-cutoff 0.05 \
       --qvalue 0.01 \
       --name sample_name
```

Otherwise, it is possible for MACS2 not to build a model `--nomodel`, shift the ends toward the 3′->5′ direction, `--shift -100`, and extend the reads in the 5′–>3′

direction, `--extsize 200`. This enables analysis of the position of Tn5 cut sites rather than analysis of whole fragments, which can span nucleosomes.

```
macs2 callpeak \
       --treatment sample.bam \
       --gsize mm \
       --nomodel \
       --shift -100 \
       --extsize 200 \
       --broad \
       --broad-cutoff 0.05 \
       --qvalue 0.01 \
       --name sample_name
```

By default, MACS2 de-duplicates data by keeping a maximum of one read that has the same coordinates on the same strand (`--keep-dup 1`). To prevent this, use `--keep-dup all` to keep all reads, or `--keep-dup auto` to let MACS2 decide, based on binomial distribution using 1e-5 as a *p* value cutoff, the maximum number of reads to keep.

Genome-wide analysis of differential enrichment/accessibility

Differential analysis of accessibility/enrichment requires the generation of a consensus feature set containing the union of all called peaks among the samples being compared, usually filtered on those peaks that appear in two or more replicates. The read density in each feature of the consensus feature set is then calculated for each of the original samples. These counts are then fed into differential analysis algorithms such as DESeq2 or edgeR for standard normalization and differential analysis. A package such as DiffBind can be used to process the samples through the steps above. To run DiffBind, it is necessary to provide a list of samples along with their files names and the following information within a `csv` file:

```
SampleID Tissue Treatment   Condition Replicate bamReads
ControlID bamControl    Peaks    PeakCaller
```

The first step is to read the peak sets defined in the sample list to create a DBA object, followed by the calculation of reads density in each feature. During this operation, the option `minOverlap=X` instructs DiffBind to include peaks only if they are present in X replicate samples (we recommend using a value of 2), while the `score` defines which measurement of read count to use in the binding affinity matrix (we recommend `DBA_SCORE_RPKM`)

```
library(DiffBind)

dba <- dba(sampleSheet=“samples.csv”)

dba <- dba.count(atacDBA.csv, minOverlap=2, score=DBA_
SCORE_RPKM)
```

To compare samples (referred to as a contrast in Diffbind), the `categories` option is used to define which of the columns in the sample list file to base the contrast on. For differential analysis, it is possible to use both edgeR and DESeq2 to analyze differential binding by passing either to the `method` option. We generally use the option `bFullLibrarySize=TRUE` to enable use of the total number of reads in each sample (full library size), rather than the total number of reads overlying the consensus peak set in each sample, which is recommended if overall binding levels are different among samples:

```
dba <- dba.contrast(dba, categories=DBA_FACTOR)

Dba <- dba.analyze(dba, bFullLibrarySize=TRUE,method=c
(DBA_EDGER,DBA_DESEQ2))
```

Once a set of contrasts have been established, functions within the diffbind package can be utilized to visualize data, such as dba.plotMA and dba.plotVenn, or to perform basic logical and numerical operations to determine uniquely and differentially enriched loci. However, we recommend outputting GRanges objects containing differential peak sets for further specific analysis and graphical representation within the general R environment.

Motif calling

To identify the consensus DNA sequence motifs enriched in a set of called peaks or a set of differentially enriched peaks between a set of experimental samples, a motif discovery tool such as HOMER or MEME-suite can be used. Such tools can be used to define both de novo consensus motifs enriched in the specified peak set and known motifs that are drawn from motif databases such as JASPAR. The following command is used to run motif analysis using HOMER on genomic regions (e.g., the peaks called by MACS2):

```
findMotifsGenome.pl sample.bed GENOME sample.motif -len 8,
10,12 -size 200 -preparsedDir preparsedDir
```

where `sample.bed` provides the set of genomic intervals to be analyzed, `GENOME` is the reference genome (e.g., mm9), and `-size` defines the size of the regions being analyzed (default: 200). In addition, `-len` is used to define the lengths of motifs searched by HOMER. It is possible to change these lengths using the `-len <#,#,#>` format with no spaces between the numbers.

Nucleosome positioning using ATAC-seq data

ATAC-seq can also be used to determine nucleosome positioning surrounding accessible chromatin (Buenrostro et al., 2013). Nucleoatac is a suite of tools that can be used to facilitate such analysis. It requires paired-end BAM files, a reference genome as fasta file, and broad open chromatin regions as a bed file. It is recommended to extend these genomic regions (e.g., by 200 bp) and to merge overlapped ones using bedtools as follows:

```
cat broad_regions.bed \
| bedtools slop -b 200 -g genome.chrom.sizes \
| sort -k1,1 -k2,2n \
| bedtools merge > broad_regions.slop.bed
```

To run nucleoatac, use the following commands.

```
nucleoatac run \
    --bed broad_regions.slop.bed \
    --bam sample.bam \
    --fasta genome.fasta \
    --out sample.output \
    --cores 2
```

In summary, the above steps enable use of a peak caller (MACS2) to identify regions with enriched reads (peaks) in each sequencing sample. Next, using DiffBind, the enriched regions were tested for their differential enrichment or accessibility among different samples. We used HOMER to search for known and de novo motifs to identify plausible binding of TFs. For ATAC-seq, the nucleosome positioning around the enriched accessibility regions was determined using NucleoATAC.

REAGENTS AND SOLUTIONS

ATAC-seq lysis buffer

To 980 µl of RNase/DNase-free water add the following to make 1 ml 1× ATAC-seq lysis buffer:

10 µl 1 M Tris·Cl, pH 7.4 (Current Protocols, 2001)
2 µl 5 M NaCl
3 µl 1 M $MgCl_2$
5 µl 20% (v/v) Igepal CA-630
Stored at 4°C up to 1 week

EDTA/SDS, 20×

Add 2 ml of 0.5 M EDTA and 10 ml 10% (v/v) SDS to 38 ml of RNase/DNase-free water. Store indefinitely at room temperature.

LiCl buffer

To prepare 500 ml of LiCl buffer, add to 372.5 ml RNase/DNase-free water the following:

125 ml 1 M LiCl
2.5 ml 100% NP-40
2.5 g sodium deoxycholate
Store indefinitely at room temperature

RIPA buffer

To prepare 1000 ml of RIPA buffer, add to 968 ml RNase/DNase-free water the following:

10 ml 1 M Tris·Cl, pH 7.6 (Current Protocols, 2001)
2 ml 0.5 M ethylenediaminetetraacetic acid (EDTA)
10 ml 10% (v/v) SDS
1 g (w/v) sodium deoxycholate
10 ml Triton X-100
Store indefinitely at room temperature

RIPA + 0.3 M NaCl

Add 1.8 ml 5 M NaCl to 30 ml of RIPA buffer (see recipe). Store indefinitely at room temperature.

Shearing buffer

Add 12.5 ml 1 M Tris·Cl, pH 7.6 (Current Protocols, 2001) and 0.5 ml Triton X-100 to 237 ml RNase/DNase-free water. Store indefinitely at room temperature.

Shearing buffer + PI

Add 1 tablet of complete protease inhibitors (Roche) and 100 µl of phenylmethylsulfonyl fluoride (PMSF) to 10 ml of shearing buffer and vortex until the tablet is dissolved. Use within 24 hr.

TE buffer, pH 8.0 + 0.2% Triton X-100

Add 1 ml of Triton X-100 to 500 ml of TE, pH 8.0
Can be stored at room temperature long term.

COMMENTARY

Background Information

ChIP-seq is used to measure the genome-wide distribution of specific proteins and to detect post-translational histone modifications. To measure the distribution of accessible chromatin, several techniques have been developed such as DNase-seq, MNase-seq, and ATAC-seq. ATAC-seq has become the preferred method in many labs since it requires low input amounts and provides a fast and sensitive method to map accessible DNA loci.

Critical Parameters and Troubleshooting

Selection

Size selection is an important step to improve sequencing efficiency. On the Illumina platform, small DNA fragments tend to cluster more efficiently, and they out-compete the larger ones. Additionally, shorter DNA fragments may reduce sequencing capacity by reading through to the opposite adaptor and generating overlaps in paired-end libraries. Finally, shorter DNA fragments may contain undesired reads such as adapter-dimers or primer-dimers.

Library sequencing

In a single-end run, the DNA fragment is sequenced from one end only, while in a paired-end (PE) run, the DNA fragment is sequenced additionally from the opposite end, producing two reads (or read mates): referred to as forward and reverse reads. Paired-end sequencing improves the accuracy of aligning the read to the genome, since a longer read sequence is available for mapping. In particular, PE sequencing is useful to enhance the mappability of repetitive regions of the genome. In addition, paired-end sequencing reduces the chance that a natural duplicate read will be misinterpreted as a duplicate using fragment position–based de-duplication. In addition, paired-end sequencing is required for nucleosome positioning to be determined using ATAC-seq.

Cross-linking (ChIP-seq)

It is important to practice accurate timing during the cross-linking step. Longer exposure to formaldehyde results in unspecific DNA-protein and protein−protein bindings, while shorter exposure may lead to loss of specific interactions.

Blacklisted regions (ChIP-seq)

Multi-mapping reads that map to genomic regions with anomalous, unstructured and repetitive regions (e.g., centromere, telomeres, etc.) could increase the level of false-positives called peaks. Therefore, it is recommended to exclude these regions from BAM files before peak calling.

Avoiding tagmentation of DNA released by apoptotic cells (ATAC-seq)

Once activated, lymphocytes can undergo high levels of apoptosis. Therefore, when preparing lymphocytes for ATAC-seq, it is recommended to FACS sort live cells prior to the preparation of nuclei for tagmentation. This prevents undesired reads that result from apoptotic DNA from leading to a high level of false-positive peaks.

Mitochondrial reads (ATAC-seq)

A high percentage of reads (40% to 60%) map to the mitochondrial genome, and these reads should be removed either manually or using blacklisted regions (Montefiori et al., 2017). Another suggestion is to include a step in the ATAC-seq protocol where targeted CRISPR/Cas9 is used for cleavage of mitochondrial DNA fragments.

Understanding Results

To visualize sequenced data, genome browser software can be used, such as The Integrative Genomics Viewer (IGV; Robinson et al., 2011). Due to the large size of the aligned BAM files, it is recommended to convert these files into tdf format (`igvtools count file.sorted.bam file.sorted.tdf mm9`). Figure 2A shows four genomic tracks representing a ChIP-seq sample, an input sample, and two ATAC-seq samples. Although the ChIP-seq track shows distinct enriched peaks in this genomic locus, the input track shows evenly distributed low signal, which represents the background. ATAC-seq peaks show the difference in accessibility in this locus between the ATAC1 (more accessible) and ATAC2 (less accessible) samples. The motif finder, HOMER, lists the enriched sequences or motifs identified

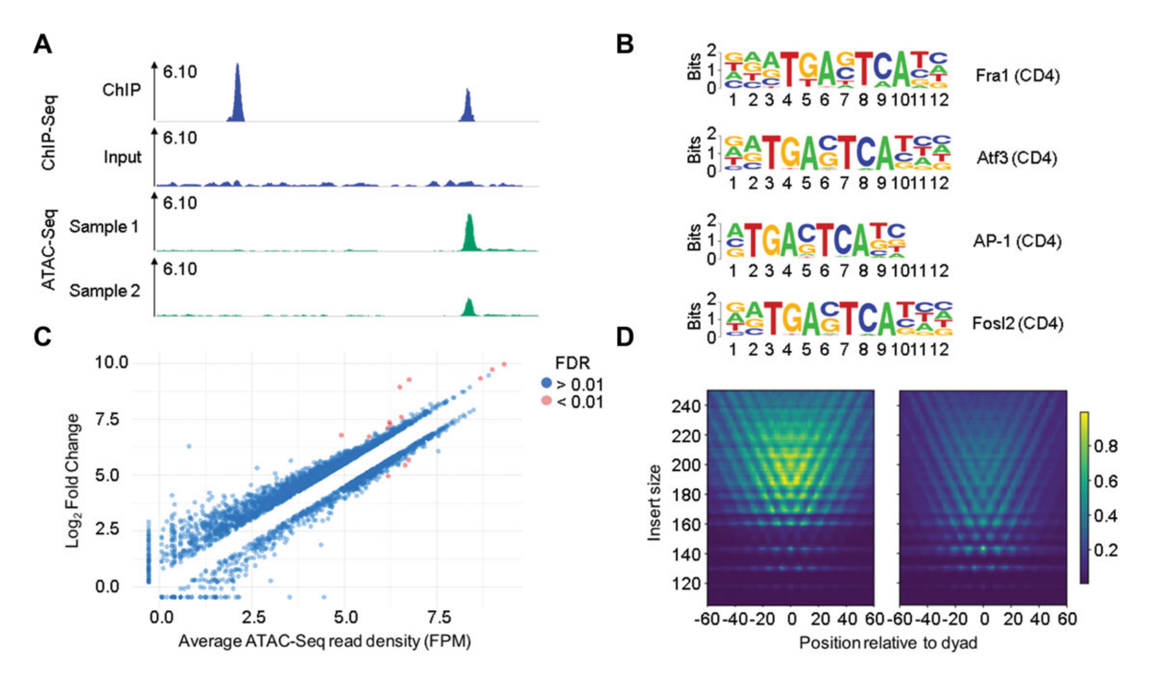

Figure 2 Example graphical outputs generated using the Basic Protocol 4 workflow. (**A**) Alignments showing read density following sequencing of corresponding ChIP and input samples (top two traces) and chromatin accessibility in indicated ATAC-seq samples. (**B**) Enriched motifs in differentially accessible peaks between two ATAC-seq experimental groups identified by Diffbind using HOMER. (**C**) MA-plot representing the average expression and fold change of global called ATAC-seq peaks between two experimental groups. Each dot represents a called peak in the union peakset. (**D**) V-plot showing the density of fragment sizes versus fragment center position (e.g., nucleosome dyad positions).

in called peaks (Fig. 2B). These motifs represent different binding sites of the protein of interest (ChIP-seq) or the proteins involved in increased or decreased accessibility (ATAC-seq). The R package DiffBind is used to detect differential binding or accessibility, and outputs the results using different plots. One important plot is the MA plot, which graphically represents fold change and mean expression (Fig. 2B). The Python package, NucleoATAC, is used to characterize the ATAC-seq signal around nucleosome positions. One of the plots produced is a V-plot, which maps the density of fragment sizes versus fragment center position (e.g., nucleosome dyad positions). Here the apex of the V-shape represents the smallest possible fragments that span the DNA protected by a nucleosome. The differences in V-pattern between samples I and II represent steric hindrance of the transposase by nucleosomes (Fig. 2D).

Acknowledgments

The research was supported by a Wellcome Trust/Royal Society grant 105663/Z/14/Z, the UK Biotechnology and Biological Sciences Research Council grants BB/N007794/1, BBS/E/B/000C0427, and BBS/E/B/000C0428, and Cancer Research UK grant C52623/A22597. The authors thank Francis Grant and Simon Andrews for their advice and discussion.

Literature Cited

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., . . . Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837. doi: 10.1016/j.cell.2007.05.009.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., . . . Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, *132*(2), 311–322. doi: 10.1016/j.cell.2007.12.014.

Boyle, A. P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., . . . Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, *21*(3), 456–464. doi: 10.1101/gr.112656.110.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. doi: 10.1038/nmeth.2688.

Current Protocols. (2001). Common media, buffers, and stock solutions. *Current Protocols in Immunology*, *34*, A.2A.1–A.2A.8. doi: 10.1002/0471142735.ima02as34.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., … Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, *6*(4), 283–289. doi: 10.1038/nmeth.1313.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*(5830), 1497–1502. doi: 10.1126/science.1141319.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., … Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813–1831. doi: 10.1101/gr.136184.111.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, *9*(4), 357–359. doi: 10.1038/nmeth.1923.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., … Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, *448*(7153), 553–560. doi: 10.1038/nature06008.

Montefiori, L., Hernandez, L., Zhang, Z., Gilad, Y., Ober, C., Crawford, G., … Jo Sakabe, N. (2017). Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Scientific Reports*, *7*(1), 2451. doi: 10.1038/s41598-017-02547-w.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., … Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, *489*(7414), 83–90. doi: 10.1038/nature11212.

Ponts, N., Harris, E. Y., Prudhomme, J., Wick, I., Eckhardt-Ludka, C., Hicks, G. R., … Le Roch, K. G. (2010). Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Research*, *20*(2), 228–238. doi: 10.1101/gr.101063.109.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. doi: 10.1093/bioinformatics/btq033.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., … Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, *4*(8), 651–657. doi: 10.1038/nmeth1068.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., … Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*, 24–26. doi: 10.1038/nbt.1754.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., … Carroll, J. S. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, *481*(7381), 389–393. doi: 10.1038/nature10730.

Schep, A. N., Buenrostro, J. D., Denny, S. K., Schwartz, K., Sherlock, G., & Greenleaf, W. J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Research*, *25*(11), 1757–1770. doi: 10.1101/gr.192294.115.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., … Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, *132*(5), 887–898. doi: 10.1016/j.cell.2008.02.022.

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, *31*(12), 2032–2034. doi: 10.1093/bioinformatics/btv098.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., … Liu, X. S. (2008). Model-based analysis of ChIP-seq (MACS). *Genome Biology*, *9*(9), R137. doi: 10.1186/gb-2008-9-9-r137.