Chapter 8

Identification of Transcribed Enhancers by Genome-Wide Chromatin Immunoprecipitation Sequencing

Steven Blinka, Michael H. Reimer Jr., Kirthi Pulakanti, Luca Pinello, Guo-Cheng Yuan, and Sridhar Rao

Abstract

Recent work has shown that RNA polymerase II-mediated transcription at distal *cis*-regulatory elements serves as a mark of highly active enhancers. Production of noncoding RNAs at enhancers, termed eRNAs, correlates with higher expression of genes that the enhancer interacts with; hence, eRNAs provide a new tool to model gene activity in normal and disease tissues. Moreover, this unique class of noncoding RNA has diverse roles in transcriptional regulation. Transcribed enhancers can be identified by a common signature of epigenetic marks by overlaying a series of genome-wide chromatin immunoprecipitation and RNA sequencing datasets. A computational approach to filter non-enhancer elements and other classes of noncoding RNAs is essential to not cloud downstream analysis. Here we present a protocol that combines wet and dry bench methods to accurately identify transcribed enhancers genome-wide as well as an experimental procedure to validate these datasets.

Key words eRNA, Chromatin immunoprecipitation sequencing, Global run on sequencing, Noncoding RNA, Transcribed enhancer, ENCODE

1 Introduction

Enhancers are distal *cis*-regulatory elements that, in contrast to promoters, activate gene expression independent of distance and orientation. Seminal work from several groups has described a series of epigenetic marks that define enhancer elements including a combination of histone marks that predict tissue-specific enhancers and their activity [1–5]. Histone H3 lysine 4 monomethylation (H3K4me1) is a hallmark for all enhancers, whereas the presence of histone H3 lysine 27 acetylation (H3K27Ac) further defines an active enhancer [4–6]. Consistent with these observations, the COMPASS complexes (which catalyze H3K4me1) and the histone acetyltransferase p300 (which catalyzes H3K27Ac) are commonly found at active enhancers in addition to promoters and gene bodies. Genome-wide locations of enhancer elements can be identified by

Ulf Andersson Ørom (ed.), *Enhancer RNAs: Methods and Protocols*, Methods in Molecular Biology, vol. 1468,
DOI 10.1007/978-1-4939-4035-6_8,

profiling these histone marks, using chromatin immunoprecipitation coupled with next-generation sequencing (ChIP-Seq). While transcription factors, coactivators (Mediator), and low DNA methylation may be used to assist with identification enhancer elements, they are not required, thereby eliminating the need for additional cell type-specific datasets [7]. In addition, with the availability of publicly accessible databases, many of these epigenetic marks have been identified in a variety of cell types.

RNA Polymerase II (RNAPII) binds a subset of enhancers and produces a unique class of long noncoding RNAs termed eRNAs. eRNAs are bidirectionally transcribed and unspliced, making them distinct from other types of long noncoding RNAs (lncRNAs) and have been demonstrated by a number of groups to be a mark of highly active enhancers [8–14]. eRNAs have been shown to have diverse roles in regulating transcription in *cis* including stabilizing enhancer looping and regulating RNAPII phosphorylation state at gene promoters [15, 16]. Genes associated with eRNA producing enhancers are thought to be critical to controlling cell identity and lineage commitment [14]. Functionally, the enhancers described above are similar to super enhancers or stretch enhancers, which drive expression of genes critical to cell identity [17]. Identification of eRNAs is often achieved by overlaying ChIP-Seq datasets with genome-wide RNA sequencing datasets (e.g., global run on sequencing; GRO-Seq). Nonetheless, our own work demonstrates that the ChIP-Seq datasets alone can be used to identify highly active enhancers likely to produce eRNAs [14]. Rigorous analyses are essential as it is challenging to distinguish enhancer transcribed RNAs from other lncRNAs (e.g., long intergenic noncoding RNAs–lincRNAs).

Here, we describe in detail our procedure for accurate identification of eRNAs using a combination of wet and dry bench approaches. We use mouse embryonic stem cells (mESCs) as a model system because the transcriptional and chromatin regulatory networks controlling pluripotency have been well characterized on a genome-wide basis via integration of existing data sets. Specifically, we outline how to generate high quality ChIP DNA libraries for sequencing. An alternative starting point includes access to published datasets (e.g., ENCODE or GEO omnibus) that allow users to perform analyses *in silico*. Upon generation or download of ChIP-Seq datasets (H3K27Ac, H3K4me1, and RNAPII) we describe the dry bench analysis by which we: (1) define putative enhancers, (2) identify eRNA positive enhancers, and (3) exclude eRNA negative enhancers and other proximal *cis*-regulatory elements such as promoters. A dry bench strategy to eliminate non-enhancer elements (e.g., pseudogenes, microRNAs, and lncRNAs) that cloud analysis using computational approaches is essential. Lastly, we validate the ChIP-Seq data by wet bench approaches including ChIP-qPCR.

2 Materials

2.1 Solutions

1. Mouse ESC media: 500 mL Dulbecco's Modified Eagle's Medium (DMEM), 100 mL fetal bovine serum (FBS) Benchmark™, 12.5 mL Penicillin-Streptomycin Solution 100×, 6.25 mL l-glutamine, 100× liquid, 6.25 mL MEM non-essential amino acids, 6.25 mL EmbryoMax® Nucleosides (100×), 4.4 μL 100% 2-mercaptoethanol, 62.5 μL leukemia inhibitory factor (LIF). The final concentration of LIF is 10^3 μ/mL. Good for 3–4 weeks at 4 °C.
2. 1× DPBS: 5 mL 10× DPBS, 45 mL autoclaved reverse osmosis (RO) water.
3. 2.5 M glycine: 187 g glycine, 1 L RO water. Good for 1 year.
4. 70% Ethanol: 35 mL 100% ethanol, 15 mL RO water.
5. 10% sodium deoxycholate: 5 g sodium deoxycholate, 50 mL RO water.
6. SDS Lysis Buffer: 250 μL 20% SDS, 200 μL 0.5 M EDTA, 1.5 mL 5 M NaCl, 500 μL Triton X-100, 1 mL Tris–HCl pH 8.0, 46.6 mL RO water. Store at 4 °C. Good for 6 months when stored without protease inhibitors.
7. Low-Salt Wash Buffer II: 250 μL 20% SDS, 200 μL 0.5 M EDTA, 1.5 mL 5 M NaCl, 500 μL Triton X-100, 1 mL Tris–HCl pH 8.0, 46.6 mL RO water. Good for 6 months.
8. Wash Buffer III (LiCl): 2.5 mL 5 M LiCl, 2.5 mL 10% NP40, 2.5 mL 10% deoxycholate, 100 μL 0.5 M EDTA, 500 μL Tris–HCl pH 8.0, 41.9 mL RO water. Good for 6 months.
9. TE: 500 μL 1 M Tris–HCl pH 8.0, 100 μL 0.5 M EDTA, 49.4 mL RO water.
10. SDS Elution Buffer: 2.5 mL 20% SDS, 1 mL 0.5 M EDTA, 2.5 mL 1 M Tris–HCl, 44 mL RO water. Good for 6 months.

2.2 Lab Equipment

1. Phase lock gel tubes—heavy (5 PRIME—2302810).
2. Qubit® Fluorometer.
3. Dynal magnetic separation rack.
4. Qsonica Q125 Sonicator with 1/8″ in diameter tip or Diagenode Bioruptor® Pico.
5. 1.5 mL Bioruptor® microtubes (Diagenode C30010016) if using the Bioruptor® Pico.
6. Eppendorf Tubes® 5.0 mL if using the Qsonica Sonicator.
7. Bioanalyzer.
8. AMPure® XP Beads.

2.3 Chemicals

1. Mouse ESC media.
2. 1× DPBS.

3. 2.5 M glycine.
4. 70% Ethanol.
5. 10% Sodium deoxycholate.
6. SDS Lysis Buffer.
7. Low-Salt Wash Buffer II.
8. Wash Buffer III (LiCl).
9. TE.
10. SDS Elution Buffer.
11. 16% Methanol-free formaldehyde.
12. Protease inhibitor cocktail.
13. Phenylmethylsulfonyl fluoride (PMSF).
14. Protein A or G beads.
15. Phenol:Chlorofrom:Isoamyl Alcohol (25:24:1 v/v).
16. Agarose powder.
17. RNase A.
18. Proteinase K.
19. Glycogen.
20. 3 M Sodium acetate.

2.4 Kits

1. Qubit® dsDNA HS Assay Kit.
2. NEBNext® ChIP-Seq Library Prep Master Mix.
3. NEBNext® Singleplex or Multiplex Adapters. Adapter combinations will vary based on sample number and complexity of library. Refer to protocol for pooling and adapter ligation included with the ChIP-Seq library kit.

2.5 ENCODE Antibodies

1. H3K4me1 (Abcam ab8895).
2. H3K27Ac (Abcam ab4729).
3. RNA Polymerase II (Abcam 8WG16).
4. H3K36me3 (Abcam ab9050-optional).

3 Methods

3.1 Wet Lab Protocol for ChIP

3.1.1 Cell Preparation and Chromatin Immunoprecipitation

This protocol is designed to perform ChIP in mESCs for endogenous proteins and will need to be optimized for additional cell types. The total time from starting the procedure to having ChIP DNA ready for downstream processing (such as quantitative PCR or ChIP-Seq library generation) is 4 days (Fig. 1). This does not include the preparation/splitting of cells [18–22].

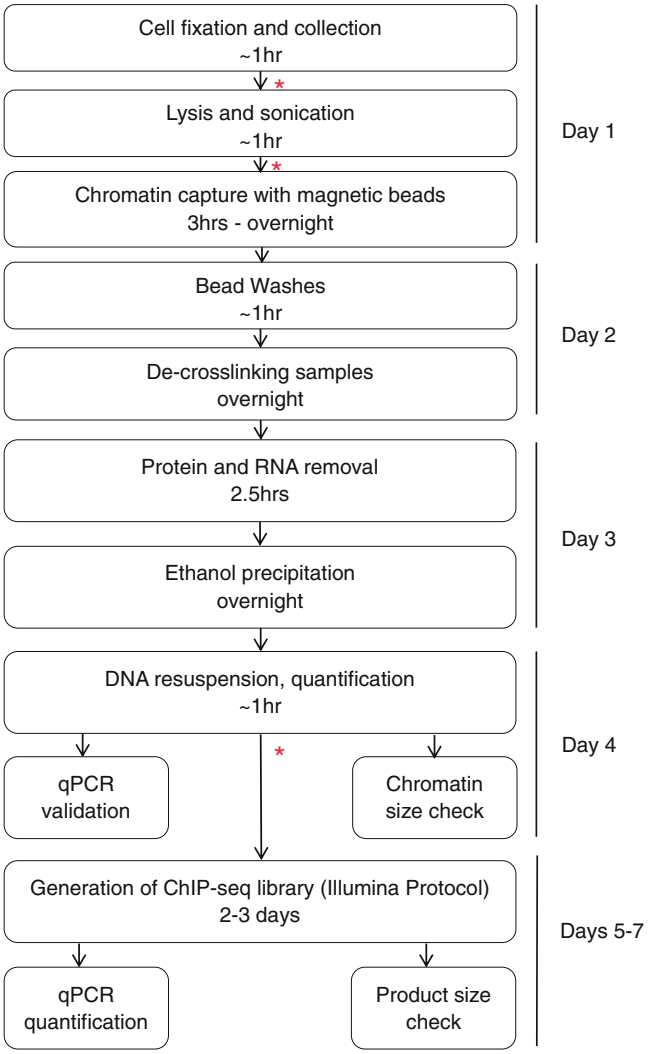

Fig. 1 Flowchart of wet bench protocol to generate ChIP-Seq library. * Indicates a safe stopping point in the protocol, overnight or a couple days

Prior to Day 1

Prepare mESCs on gelatin-adapted plates. This protocol is written for two 15 cm plates that are 30–60% confluent, which would yield approximately 100 million cells total (*see* **Note 1**).

Day 1

1. Add 1250 μL of 16% formaldehyde to 20 mL media and cells (final concentration 1.0%). Incubate at room temperature for 5 min with gentle rocking to mix (*see* **Note 2**).
2. Quench formaldehyde with 1.0 mL of 2.5 M Glycine (final concentration 125 mM). Incubate at room temperature for 5 min with gentle rocking to mix.
3. Rinse plate with 20 mL ice-cold Dulbecco's phosphate-buffered saline (DPBS) (without magnesium and calcium)

containing 1:1000 protease inhibitors (PI) and 1:200 phenylmethylsulfonyl fluoride (PMSF). Rinse fixed cells 3× total. Keep plates on ice while rinsing fixed cells.

4. Add 15 mL DPBS containing inhibitor 1:100 PI and 1:200 PMSF and scrape fixed cells into a 50 mL conical tube on ice. Rinse plate two more times with 15 mL DPBS and collect with initial scraping.
5. Centrifuge at 750 × *g* for 10 min at 4 °C to and aspirate supernatant. Transfer cell pellet to a 1.7 mL microcentrifuge tube and flash freeze the sheared chromatin pellet on Dry Ice. Store at −80 °C for up to several months.
6. Lyse cells with 1 mL SDS Lysis Buffer containing inhibitors (1:100 PI and 1:200 PMSF) for each 15 cm plate that was approximately 30–60 % confluent to start. Pipette up and down to break apart aggregates of fixed cells.
7. Transfer to a 5 mL Eppendorf tube and incubate 10 min on ice. A large 5 mL tube allows the Qsonica microtip to be inserted without touching the sides of the tube, yet still come very close to the bottom of the conical (*see* **Note 3**).
8. Proceed with sonication using a microtip. Each sample should receive three cycles at Amplitude = 5 in ice water. Each cycle should consist of a burst of 1 s on and 4 s off, for a total of 30 s on. There should be a 3 min pause between each cycle (*see* **Notes 4** and **5**).
9. Pellet insoluble fraction by spinning at maximum speed for 10 min at 4 °C. Transfer supernatant to a new 1.7 mL microcentrifuge tube.
10. Remove a small aliquot (100 µL) to be saved as Input/genomic DNA in a screw cap microcentrifuge tube. Store at −80 °C. If needed, the samples can be frozen at −80 °C for months.
11. Boil 50 µL of each sample for 15 min. Spin at max speed in a microcentrifuge for 5 min at room temperature. Run 10–20 µL on a 1% agarose gel. The bulk of the decross-linked DNA should be 200–500 base pairs (bp) (*see* **Note 6**) (Fig. 2).
12. Add 4–8 micrograms of antibody to chromatin and place at 4 °C overnight (*see* **Note 7**).

Day 2

1. Pipette 50–100 µL of Protein A or G Dynabeads into a fresh 1.7 mL microcentrifuge tube and place into magnetic separation rack for 2 min (*see* **Note 8**).
2. Remove liquid using a 1 mL micropipette. Resuspend in 1 mL ChIP Lysis Buffer with 1:1000 PI and 1:200 PMSF and rotate for 5 min at 4 °C.
3. Quick spin samples to pull down liquid from cap and place tubes into magnetic separation rack for 2 min and remove liquid. Wash beads 3× total in ChIP Lysis Buffer with inhibitors.

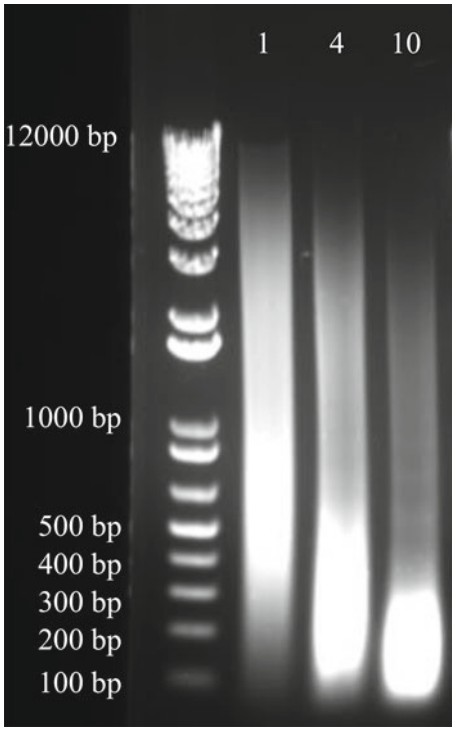

Fig. 2 Approximately 20 million mouse embryonic stem cells were fixed for 5 min with 1 % formaldehyde. 300 μL of each sample was sheared in 0.1 % SDS Lysis Buffer using the Diagenode Bioruptor® Pico. From left to right on the gel, samples were subjected to 1, 4, and 10 cycles of sonication. One microgram of decross-linked and RNase/proteinase K sample was separated by electrophoresis on a 1 % agarose gel that was stained with ethidium bromide. Optimally, sheared chromatin will yield a smear between 200 and 500 bp (as seen with 4 cycles above). One cycle yields under-sheared chromatin (400–1000+ bp) and ten cycles produces over-sheared chromatin (100–300 bp)

4. Transfer supernatant containing sheared chromatin and antibody to tubes with washed Dynabeads.
5. Rotate for at least 3 h at 4 °C.
6. Quick spin the samples to bring down the liquid and place in magnetic rack for 2 min.
7. Remove liquid with a 1 mL micropipette to avoid disturbing beads.
8. Washes can be performed at room temperature and should be quick to prevent the beads from drying out.
9. Wash the tubes using the following procedure: add 1 mL of wash buffer and resuspend by pipetting, place in tube rotator at 4 °C for 10 min, quick spin to bring down the liquid, place in Magnetic Rack for 2–3 min at room temperature, carefully

pipette all liquid without disturbing beads, remove sample from rack, and proceed with next wash buffer. Gently pipette samples up and down to ensure aggregates of beads are broken up. You may use fresh tubes for each wash, to ensure there is no carryover.

10. Wash beads with 1 mL Buffer in the following order: ChIP Lysis Buffer (1×), Low-Salt Wash Buffer II (1×), Wash Buffer III (LiCl) (1×), and TE (1×) (*see* **Note 9**).
11. After final wash, remove all traces of TE with another spin and resuspend beads in 150 µL SDS Elution Buffer.
12. Transfer all samples to screw cap microcentrifuge tubes to minimize evaporation.
13. Incubate at 65 °C overnight (preferably in a water bath to minimize evaporation). Remove the saved Input sample and begin to process in parallel. This performs both the decross-linking and the elution in a single step.

Day 3

1. Quick spin the samples and place into magnetic rack for 3 min. Input sample should be spun at maximum speed for 10 min at room temperature. Transfer supernatant to a new microcentrifuge tube and bring volume to 200 µL with TE.
2. Place new microcentrifuge tube with supernatant into magnetic rack for another 3 min to ensure all beads are removed.
3. Add 2 µL of RNase A to each sample (including Input) and incubate at 37 °C for 30 min.
4. Add 2 µL of Glycogen and 4 µL of Proteinase K to each sample and incubate at 37 °C for 2 h. Glycogen is added as a DNA carrier.
5. Pre-spin 2 mL phase lock tubes for 2 min at maximum speed to pellet resin.
6. Transfer sample to 2 mL phase lock tube. Add 1 volume (200 µL) of Phenol:Chloroform:Isoamyl Alcohol. Mix well by inverting at least 10× and spin at maximum speed for 5 min at room temperature.
7. Add 1/10th volume (20 µL) 3 M Sodium Acetate and 2.5 volumes (0.5 mL) 100% Ethanol to the tubes. Place samples on Dry Ice until they freeze completely. Place at –20 °C overnight to maximize DNA yield.

Day 4

1. Spin at maximum speed for 15 min at 4 °C. Carefully use a 1 mL micropipette and remove supernatant, preserving the small white pellet. Quick spin a second time to ensure all liquid is at bottom of tube and remove remaining liquid.
2. Air-dry the sample for 3–5 min. Do not over-dry the DNA pellet.
3. Resuspend in 25–50 µL of water. Quantitate DNA by Qubit DNA High Sensitivity at a 1:40 dilution. Sample can now be

stored at −80 °C indefinitely for downstream applications. Aliquot samples to avoid freeze thaw (*see* **Note 10**).

4. Run a small amount of precipitated DNA from samples and Input (if you have excess DNA) in a 1 % agarose gel to ensure proper sonication (Fig. 2). Input sample is preferred for ChIP-Seq to assess enrichment of proteins.
5. ChIP DNA is quality controlled using an Agilent Bioanalyzer. The Bioanalyzer validates the size of DNA fragments (200–500 bp) and determines the concentration and purity of the sample.

3.1.2 ChIP-Seq Library Generation and Validation

1. ChIP-Seq libraries are generated using the NEBNext® ChIP-Seq Library Prep Master Mix Set for Illumina according to manufacturer's instructions (*see* **Note 11**).
2. ChIP-Seq libraries are quality controlled using an Agilent Bioanalyzer prior to sequencing. Typically, the tracing will show a narrow range of products between 150 and 500 bp depending on the size of the original ChIP DNA. Details are provided in the library generation kit to facilitate decision about whether the library is of sufficient quality to provide good quality sequencing results.
3. It is critical to validate that the ChIP-Seq library is representative of the precipitated ChIP DNA in Subheading 3.1.2.3, **step 38**. Test for enrichment of protein at active enhancers (positive control) and inactive enhancers (negative control) by ChIP-qPCR prior to sequencing. As little as 0.1 ng DNA can be used for each reaction. Perform ChIP-qPCR on the Input and include a negative control antibody (e.g., IgG sample) (*see* **Notes 12** and **13**) (Fig. 3).
4. After protein enrichment is confirmed, sequence on an Illumina HiSeq and obtain a minimum of 20–40 million reads for H3K4me1 or H3K27Ac and 10–20 million reads for RNAPII. Higher reads are used for histone marks because they typically bind larger chromatin regions rather than a specific DNA element. Paired-end sequencing can be performed, but typically does not provide additional information. Indexing will depend on the run type and number of samples.

3.2 Dry Bench Analysis of ChIP-Seq to Identify Putative Enhancers

Discriminative filters and thresholds are used to specifically identify enhancers and not other *cis*-regulatory elements that may act as distal or alternative promoters. Moreover, non-enhancer elements that produce other classes of long noncoding RNAs need to be eliminated so they do not cloud analyses and computational approaches. Additionally, intragenic enhancers must be filtered to eliminate coding strand transcripts. For optimum computational performance, a minimum system requirement of 4 cores and 16GB RAM or more is recommended. Most of the ChIP-Seq computational analyses are done in Unix-like operating systems given the availability of several

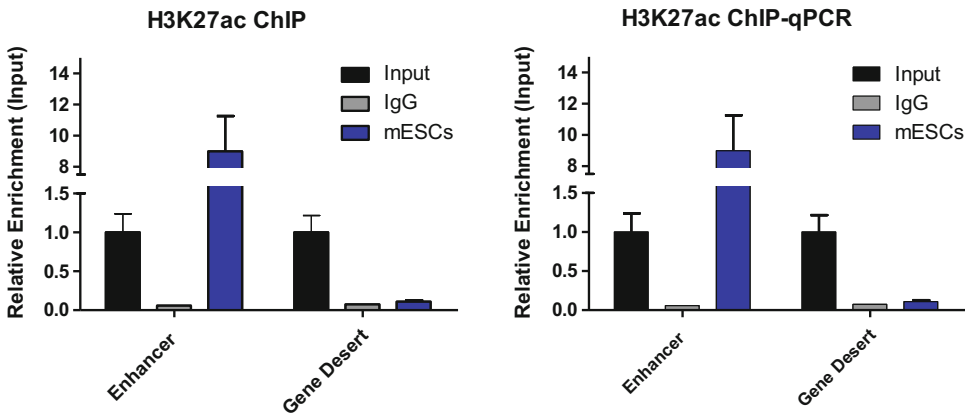

Fig. 3 ChIP-qPCR showing enrichment of the activating histone mark H3K27Ac at a pluripotency associated enhancer. Primers within a gene desert on chromosome 6 were used as a negative control. Rabbit IgG was used as mock control. Values were normalized to primers within the promoter of GAPDH

methods targeting these systems and that often provide a command line interface for their execution. R and Python are used to perform statistical analysis and to automatize analysis of the genomic data. In this section, we describe the tools and data formats used to analyze ChIP-Seq datasets to define putative enhancers (Fig. 4). The dry bench datasets generate a variety of different types. For a brief overview of file types and the data they contain, please see http://www.broadinstitute.org/software/igv/FileFormats

3.2.1 Data Mining and Retrieval

Public ChIP-Seq Data Files

1. If ChIP-Seq datasets are published for your tissue of interest, they can be downloaded from a freely available online repository such as GEO Omnibus (http://www.ncbi.nlm.nih.gov/geo/), RIKEN-FANTOM (http://fantom.gsc.riken.jp/data/), or EMBL-EBI (http://www.ebi.ac.uk/ena) [23].
2. Use the SRA toolkit that has a set of data-dump utilities, which will allow reformatting from SRA to FASTA, FASTQ, or SAM. The SRA toolkit can be downloaded from NCBI (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) and is available for Mac, Linux, and Windows operating systems.
3. Use "fastq-dump" utility to convert .sra to .fastq file format to generate a FASTQ file from SRA file(s). Each read/sequence in FASTQ file consists of 4 lines. The first line starting with "@" indicates the read identifier. The second line is the actual DNA sequence. The third line starting with "+" is an optional title line. The fourth line is the quality score symbol for each base in the sequence which is encoded in ASCII character code following usually the PHRED33 convention (other quality encodings may be used depending on the Illumina software, for more information refer to: https://en.wikipedia.org/wiki/FASTQ_format#Quality).

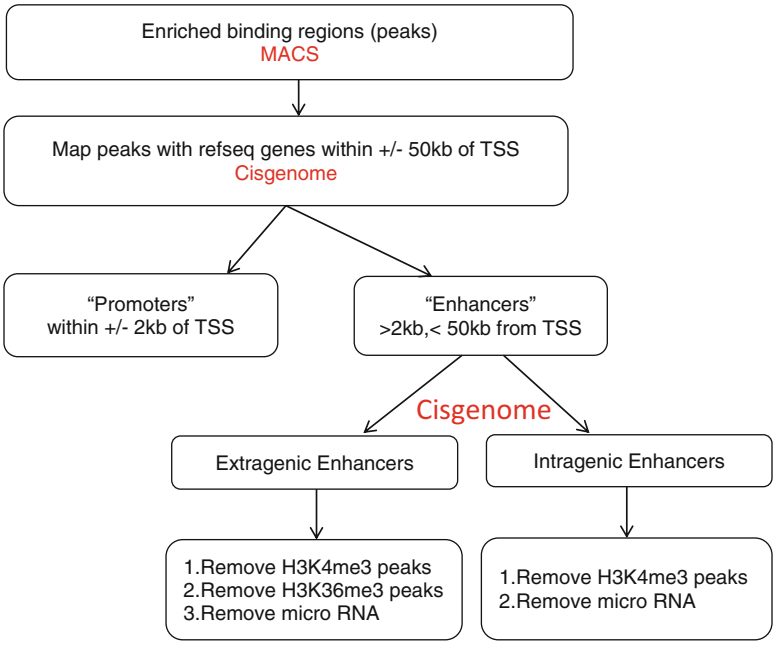

Fig. 4 Schematic representation of the workflow of enhancer detection. Peaks are called by MACS. Mapping/annotation to nearest gene is executed by Cisgenome. Distance based parsing is performed to categorize the peaks to enhancers and promoters. Enhancers are divided into extragenic and intragenic to prevent clouding of downstream analysis. Further filtration is performed to remove promoters, unannotated genes, and noncoding RNAs

Private ChIP-Seq Data from Server

1. Download raw data (typically a FASTQ file) from your sequencing instrument.
2. If data is provided in a SRA format, use the NCBI SRA toolkit to convert data to achieve a data format in order to run the alignment (*see* **Note 14**).

Quality checking FASTQ files

1. Perform quality control checks using FastQC (developed at Babraham Institute) to ensure that raw data from single-end reads is free of biases (originating from sequencing or library preparation) [24]. Decoding quality scores in a FASTQ file depends on the type of platform used. Sanger, Illumina, Solexa, and PHRED software reads the DNA sequences, calls bases, and assigns a quality value for each base called. PHRED33 quality score is the most common quality metric adopted. For Illumina the quality ranges from 0 to 62 and base pair quality score of 20 is minimally required to trust the DNA nucleotide identified. Some modules in FASTQC that are helpful to judge your sequence are Per Base Sequence Quality Report which can help you decide if sequence trimming is needed before alignment. The Sequence Duplication Level Report is informative for library enrichment. The Overrepresented Sequence Report assesses for adapter contamination.

2. Check for barcodes after downloading the FASTQ file. Any adapter sequences that are used in sequencing library construction should be trimmed, for example, using the Trimmomatic utility (http://www.usadellab.org/cms/?page=trimmomatic) [25].
3. Trim low quality sequences. Low quality reads could have high sequencing error, resulting in misalignment to the reference genome. Other preprocessing tasks for FASTQ files such as filtering sequences based on quality, formatting the width of sequences, converting the FASTA sequence to RNA/DNA, etc. can be done using FASTX-toolkit [26].

Mapping Reads to the Reference Genome

4. ChIP sequencing is most often performed with single-end reads. Use Bowtie 1.1.2 algorithm to align single-end reads to mouse genome build mm9 using parameters (–p 6 –n 2 –l 49 –e 70 –m 1 – – best for unique mapping), which allows a maximum of 2 mismatches (n) in the 49 bases (l) and 1 unique alignment per read (m) and uses 6 cores (p). If your machine has more than 6 cores you should adjust this parameter (*see* **Note 15**) [27].
5. The output is a TAB-delimited Sequence Alignment/MAP (SAM) file describing mapped alignments (now known as "tags") of sequencing reads to a reference sequence.
6. Convert SAM to BAM format using Samtools [28, 29]. BAM is a compressed and binary equivalent of SAM.
7. Use SAMtools which has a set of utilities to manipulate the alignments in BAM format for further downstream analysis. SAMtools can be adopted for merging, sorting, and indexing the BAM files.

3.2.2 Peak Calling

1. Peak calling is done to identify the binding sites for RNAPII or histone modifications. MACS2 2.1.0 (model-based analysis of ChIP-Seq) is used to identify significantly enriched regions (sites of DNA-protein binding peaks) over background (the Input samples used to estimate the per base pair noise levels) [30].
2. MACS reports all binding sites with p-values below a defined threshold (default 10^{-5}) in a BED format. Set a p-value threshold of enrichment of 10^{-5} and option – – broad to identify H3K4me1 and H3K27Ac ChIP-Seq data since distribution of histone reads have a continuous property and peaks are broad. Only H3K4me1 and H3K27Ac peaks greater than 1 kb in length are considered in this analysis. This assists with eliminating spurious genomic regions that are less likely to possess enhancer function. In addition, given that eRNAs are a lncRNA, this size discrimination assists in eliminating other elements that may produce small noncoding RNAs. A p-value of 10^{-6} is used to detect narrow well-defined (non-broad) RNAPII ChIP-Seq data (*see* **Note 16**).

3.2.3 Putative Enhancer Detection

Enhancers are noncoding DNA elements that act independent of distance and orientation to regulate gene transcription. However, many of the marks described above are not exclusive to enhancers. As a result, genomic elements that mimic transcribed enhancers (e.g., pseudogenes and microRNAs) must be removed to allow for more accurate analysis of eRNA producing enhancers. However, more sophisticated analyses require generation or availability of additional histone ChIP-Seq datasets.

1. Map the called peaks in the previous step with the nearest genes using UCSC RefFlat annotations. Use Cisgenome tool to map all ChIP-Seq tag peaks to annotated genes that are ±50 kb of TSS [31]. Specifically, use the feature refgene_getnearestgene with options –r 1 –up 50,000 –down 50,000.
2. Remove any peaks located in promoter regions. For this analysis, promoters are defined as regions 2 kb upstream and downstream of the TSS (4 kb total) (*see* **Note 17**).
3. The resulting peak list should have all the enhancer regions between 2 and 50 kb of the nearest neighbor gene TSS. Enhancers >50 kb from the TSS of a gene can be saved by altering the options in 3.2.3.1, if desired.
4. To determine whether enhancers are located within actively transcribed genes, use Cisgenome (refgene_getlocationsummary) to classify the enhancer as intragenic versus extragenic.
5. This step requires additional histone modification datasets. Use BEDTools to eliminate extragenic and intragenic enhancers that overlap with a region of H3K4me3 to remove any unannotated gene or other classes of ncRNAs. Extragenic enhancers that overlap with H3K36me3 regions should be eliminated for the same reason. This cannot be used for intragenic peaks since many intronic and exonic enhancers may show some degree of H3K36me3 enrichment (*see* **Notes 18** and **19**) [32].

3.3 Validating ChIP-Seq Data

3.3.1 Detection of Transcribed Enhancers by GRO-Seq Overlay

There is rapidly growing evidence that eRNA production is a mark of a highly active enhancer and that eRNAs have diverse roles transcriptional regulation. eRNA producing enhancers can be identified by overlapping RNAPII bound enhancers with GRO-Seq datasets. Not surprisingly, enhancers bound by RNAPII show higher eRNA production rates than unbound sites (*see* **Note 20**).

1. Use BedTools (intersectBed with –f 0.5 –r) to identify enhancers that overlap with RNA Pol II (50% minimum overlap). We have found that enhancers occupied by RNAPII are highly enriched for eRNA production (*see* **Note 21**) [14].
2. To estimate expression levels for enhancers that are bound by RNAPII, processed GRO-Seq data available on GEO omnibus (GSE27037) was downloaded.

3. For extragenic enhancers, use BEDTools suite (coverageBed) to count RNA reads from both strands.
4. For intragenic enhancers, use BEDTools suite (coverageBed) to count RNA from only the antisense strand to prevent counting reads from sense-strand gene transcription. Since, the sense strand is the coding strand, there should be approximately half as many reads. Accordingly, transcribed intragenic enhancers cannot be directly compared with transcribed extragenic enhancers. Genes need to be separated by coding strand and counted separately. Using this approach, intragenic enhancers that produce eRNAs can be identified, with approximately half the number of transcripts of extragenic enhancers [14].
5. Compute RPKM (reads per kilobase of genomic region per million mapped reads) for each enhancer that is associated with the nearest gene (*see* **Note 22**).

3.3.2 Wet Bench Approach to Validate Presence of H3K4me1, H3K27Ac, RNAPII, and Tissue Specific eRNAs

1. Confirm enrichment of protein at an enhancer by qPCR with ChIP DNA as described in Subheading 3.1.2.
2. Validate the presence of cell type-specific eRNA production by RT-qPCR (*see* **Note 23**).

4 Notes

1. Do not perform cross-linking on plates with a large number of dead cells. Change media the morning of cross-linking to remove dead cells. Let cells incubate for 2–3 h to ensure they equilibrate. If combining more than one plate be sure to scale up volumes.
2. Formaldehyde mediated cross-linking is one of the key aspects to both data quality and reproducibility from ChIP. Ideally, a short enough incubation time is used to cross-link DNA and proteins within close physical proximity, without causing distal interacting sites/proteins to cross-link. Fixing cells for too long may reduce the number of available epitopes and make it more difficult to lyse and shear the chromatin, thus reducing DNA yields. It may also make reverse cross-linking more difficult which will interfere with downstream steps. For the vast majority of cells, cross-linking is between 5 and 7 min at room temperature, and rarely requires more than 10 min. Use fresh formaldehyde as air and light exposure can change the contents. Methanol-free formaldehyde is preferred as methanol can disrupt cell membranes and effect lysis. We find individual ampules of methanol free formaldehyde reduces the variability from assay to assay significantly.
3. Sonication is arguably the most important step of a ChIP assay. The sonication microtip should be consistently placed as close to the bottom of the 5 mL conical tube as possible for all samples. This prevents foaming and ensures similar sonication

between samples. If there is significant frothing/foaming, pause, remove sample and spin down quickly in a microcentrifuge to remove foam, and restart. The most likely cause of frothing is because the tip is not close enough to the bottom of the tube. Make sure the microtip does not contact the tube (bottom or sides). If you see precipitate, you may want to discard the sample and fix new cells if available.

4. Each cell type requires different sonication conditions and SDS Lysis Buffer. If SDS Lysis Buffer requires a SDS concentration greater than 0.1%, samples must be diluted (final concentration of 0.1% or less) prior to adding antibody. SDS interferes with the antibody epitope interaction. It may also affect downstream PCR. Moreover detergents (e.g., SDS) can precipitate out of solution at temperatures lower than 15 °C when stored too long. Prepare fresh lysis buffer for each experiment. If using a different number of cells (by greater than a factor of 2), type of cells, tube, or sample volume, you will need to reoptimize sonication conditions to ensure adequate fragmentation in the minimal number of cycles. Optimal size fragments are in the 200–500 bp range (Fig. 2). Fragments greater than 500 bp do not pull down as well and may result in an increase in nonspecific binding in the ChIP assay. Over shearing chromatin (100 bp or less) can be detrimental to downstream applications such as ChIP-qPCR. Over shearing may also damage proteins and alter epitopes.
5. As an alternative to using a microtip, many ChIP-Seq data sets are created using a Diagenode Biorupter® for sonication. A Biorupter® allows you to shear multiple samples at one time and eliminates variation due to microtip placement. Moreover, problems noted above including frothing/foaming are eliminated. For mESCs we use sonication conditions of 30 s On, 30 s Off for 4 cycles. 1.5 mL Diagenode Bioruptor® microtubes containing 300 μL ChIP Lysis Buffer plus inhibitors with approximately 15 million cells are used for each sample.
6. Gel electrophoresis of boiled and sheared chromatin on Day 1 is a quick method to check sonication efficiency. However, to be safe, a small amount of precipitated Input/genomic DNA should be run out to confirm that the sonication was optimal. This is representative of the ChIP DNA pulled down after RNase A and Proteinase K treatment. The band range of precipitated DNA may differ from the boiled and sheared chromatin (Fig. 2).
7. When possible, use ChIP-Seq grade antibodies that are published and preferably used to create a ENCODE dataset. Using more than the indicated amount of antibody does not result in greater DNA yield and may lead to more nonspecific binding, thereby interfering with downstream analysis. Antibodies for common histone marks such as H3K4me1 result in a high yield of DNA; therefore, less sheared chromatin may be used. For more information on how to test and validate antibodies see [33].

8. To ensure magnetic bead/antibody interaction, you can use a 50:50 mixture of protein A and protein G beads.
9. For this protocol, the ChIP Lysis Buffer and Low-Salt Wash Buffer II are the same because we lyse mESCs in 0.1 % SDS. For other cell types you may have to increase the percentage of SDS in the ChIP lysis Buffer, but Low-Salt Wash Buffer II should stay at 0.1 %.
10. A fluorometry based approach is necessary to quantify ChIP DNA. Spectrometry-based methods do not distinguish between RNA, double stranded DNA, single stranded DNA, and free nucleotides. QuBit 2.0 Fluorometer is more sensitive and accurate than spectrometry-based methods because it uses a fluorescent dye that specifically intercalates into double stranded DNA. This allows quantification of very low amounts of DNA (as low as 10 pg/μL) without interference due to other nucleotide species.
11. The ChIP-Seq library preparation kit can be purchased for any platform (although Illumina HiSeq is the most common). Alternatively it is often more efficient and cost effective to have the company performing the sequencing make the library.
12. To ensure quality of the precipitated DNA always include a negative control. A good antibody for negative control in mESCs is IgG. Verify by ChIP-qPCR (Fig. 3).
13. Negative control primers can be used for all samples if designed in gene deserts. Positive control primers for RNAPII may be designed at a known active promoter or enhancer in tissue of interest. Alternatively primers can be designed after downstream ChIP-Seq analysis based on the presence of a ChIP-Seq tag peak. ChIP-Seq tag peaks can be viewed by uploading files to Integrated Genome Viewer (IGV). ChIP-Seq tag peaks correspond to enriched presence or binding of the target protein. Be sure to run a melting curve when using new primers to ensure that you are amplifying a single PCR product.
14. There are other files to browse to look for run settings, quality metrics, etc., from your sequencer report. The raw FASTQ files are necessary to publish data. Create a backup as soon as you download your raw files.
15. Bowtie2 is generally faster and more sensitive than Bowtie1 for reads longer than 50 bp. Set seed length (l) to length of the read for each data file. Specifying the number of parallel search threads (p) increases alignment throughput. Option –m and –– best in Bowtie results in fewer unique alignments than just specifying –m. For paired-ends, the alignment can be time consuming. Option –I and –X in Bowtie are critical to get fair percentage of aligned reads. Other popular short read aligner algorithms (ELAND) could be used depending upon the type

of data. BWA is used for exome sequence reads, whereas TopHat and STAR are for RNA-Seq data. A minimum of 4–10 GB of RAM is required to run Bowtie. Bowtie can utilize all cores on a node. For a single alignment run job, you can specify the number of cores to use with the –p option. The index files can be downloaded from Bowtie website for the most common assembly (mm8, mm9, mm10).

16. MACS uses control samples to minimize bias and calculates an empirical false discovery rate (FDR). p-Values vary for different datasets depending on strength of enrichment. A good way to choose the best p-value is to visualize the signal files (wiggle files) in the genome browser and to look for peaks called by MACS. To determine if the ChIP experiment worked, sort FDR from lowest to highest and then sort fold enrichment from highest to lowest and look for the number of peaks. There should be one to several thousand peaks.
17. The size of a promoter can be around 3 to 5 kb. Simple distance based calculations in Microsoft Excel were used to identify promoters in the output Cisgenome yielded. Promoters can then be removed using Microsoft Excel software.
18. Extragenic and intragenic peaks that overlap with a region of H3K4me3 (a mark of promoters) may be eliminated to remove any unannotated gene or other classes of noncoding RNAs. However, many eRNA producing enhancers have higher levels of H3K4me3; thus, this stringent filter will remove some transcribed enhancers prior to downstream analysis. Extragenic peaks that overlap with H3K36me3 (an epigenetic mark found in gene bodies and long intergenic noncoding RNAs) may be eliminated for the same reason. Intragenic peaks may not be removed since many intronic and exonic enhancers may show some degree of H3K36me3 enrichment. Elimination of these peaks can be done using BEDTools (intersectBed) [32]. The same filtering methods described in Subheading 3 can be used to identify RNAPII, transcription factors, or coactivators at enhancers.
19. IntersectBed from BedTools was used to get the overlapping regions and nonoverlapping regions.
20. GRO-Seq is more sensitive than RNA-Seq at capturing nascent RNAs, which have properties more similar to eRNAs.
21. In this protocol we describe a direct way to identify eRNA producing enhancers using RNAPII ChIP-Seq and GRO-Seq. However, transcribed enhancers can be indirectly identified by very high levels of H3K27Ac and H3K4me3.
22. Use TopHat to align GRO-Seq or RNA-Seq data to the genome and then use Cufflinks to quantify the abundance of transcript.
23. eRNAs are expressed at levels 1:100 to 1:1000 of the mRNA of the gene they are associated with. Thus, it is important to confirm detection of tissue specific expression of eRNAs and not

background/noise. By RT-qPCR we compare cDNA from pluripotent ESCs to cells that were treated with 5uM retinoic acid for 6 days which induces complete differentiation. Given that eRNAs are unspliced, it is essential to have DNA free RNA to make sure you are only amplifying cDNA (run a no reverse transcriptase control). Moreover, not all eRNAs can be converted to cDNA during reverse transcriptase using either oligo dT or random hexamers. We use BioRad iScript cDNA Synthesis Kit because it combines both oligo dT and random hexamers.

References

1. Heintzman ND, Stuart RK, Hon G et al (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39:311–318
2. Heintzman ND, Hon GC, Hawkins RD et al (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459:108–112
3. Visel A, Blow MJ, Li Z et al (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457:854–858
4. Zentner GE, Tesar PJ, Scacheri PC (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res 21:1273–1283
5. Rada-Iglesias A, Bajpai R, Swigut T et al (2011) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470:279–283
6. Creyghton MP, Cheng AW, Welstead GG et al (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107:21931–21936
7. Stadler MB, Murr R, Burger L et al (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480:490–495
8. Kim T-K, Hemberg M, Gray JM et al (2010) Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182–187
9. De Santa F, Barozzi I, Mietton F et al (2010) A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol 8, e1000384
10. Koch F, Fenouil R, Gut M et al (2011) Transcription initiation platforms and GTF recruitment at tissue specific enhancers and promoters. Nat Struct Mol Biol 18:956–963
11. Wang D, Garcia-Bassets I, Benner C et al (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474:390–394
12. Orom UA, Derrien T, Beringer M et al (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46–58
13. Lai F, Orom UA, Cesaroni M et al (2013) Activating RNAs associate with mediator to enhance chromatin architecture and transcription. Nature 494:497–501
14. Pulakanti K, Pienello L, Stelloh C et al (2013) Enhancer transcribed RNAs are produced from hypomethylated genomic regions in a Tet-dependent manner. Epigenetics 8:1303–1320
15. Schaukowitch K, Joo JY, Liu X et al (2014) Enhancer RNA facilitates NELF release from immediate early genes. Mol Cell 56:29–42
16. Maruyama A, Mimura J, Itoh K (2014) Noncoding RNA derived from the region adjacent to the human HO-1 E2 enhancer selectively regulates HO-1 gene induction by modulating Pol II binding. Nucleic Acids Res 42:13599–13614
17. Whyte WA, Orlando DA, Hnisz D et al (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153:307–319
18. Rao S, Zhen S, Roumiantsev S et al (2010) Differential roles of Sall4 isoforms in embryonic stem cell pluripotency. Mol Cell Biol 30:5364–5380
19. Kim J, Cantor AB, Orkin SH, Wang J (2009) Use of in vivo biotinylation to study protein-protein and protein-DNA interactions in mouse embryonic stem cells. Nat Protoc 4:506–517
20. Nelson JD, Denisenko O, Bomsztyk K (2006) Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nat Protoc 1:179–185
21. Broad (2010) Broad ChIP protocol for full REMC (6 marks). http://www.roadmapepigenomics.org/protocols/type/experimental/. Accessed 19 July 2015

22. Das PP, Shao Z, Beyaz S et al (2014) Distinct and combinatorial functions of Jmjd2b/Kdm4b and Jmjd2c/Kdm4c in mouse embryonic stem cell identity. Mol Cell 53:32–48
23. Barrett T, Wilhite SE, Ledoux P et al (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41:D991–D995
24. Babraham Bioinformatics (2015) FastQC. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc. Accessed 19 July 2015
25. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120
26. Hannon Lab Cold Spring Harbor (2015) FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/index.html. Accessed 19 July 2015
27. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
28. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078–2079
29. Li H (2011) Statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993
30. Zhang Y, Liu T, Meyer CA et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137
31. Ji H, Jiang H, Ma W et al (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26:1293–1300
32. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
33. Landt S, Marinov GK, Kundaje A et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813–1831