ENCODE Guidelines for eCLIP-seq Experiments

Version 1.0, December 8, 2016

Every data producer aims to generate high-quality data sets. To help achieve that goal, this document aims to provide standards and guidelines for experiments that map the genomic location of RNA binding proteins (RBPs).

Table of Contents

I) Standard Measurements for Common ENCODE Cell Types

The ENCODE Consortium has designated common cell types that will be used by all investigators. This will aid in the integration and comparison of data produced using different technologies and platforms. To ensure consistency in cell cultures prepared in different laboratories, investigators should take the measurements below.

Required Measurements and Procedures

- **Growth time/passage number.**
  For each experiment, the date at which cells were put into culture and when they were harvested should be recorded. Investigators should use the original stock after growing a culture for two months. Passage number should be assessed and recorded for primary cells. Primary cells should not exceed 6 passages. Any experiment that does not follow the officially approved protocol for that cell line/type should be noted (see metadata standards).
- **Cell density.** Cell density should be assessed for each cell culture, recorded, and submitted along with any data generated from that culture (see metadata standards).
  - The density of GM12878 cells should be maintained between 2.0 x 10^5 cells/ml and 1.0 x 10^6 cells/ml.
  - K562 cells should be grown to a maximal density of 7.5 x 10^5 cells/ml.
  - HepG2 cells should be grown to a maximum of 75% confluence.
  - HeLa-S3 should be grown to a maximal density of 5 x 10^5 cells/ml.
- **Cell cycle and gene expression state.** Gene expression experiments were performed by seven different laboratories using cells of the same lot and recommended growth conditions. Strong concordance was observed. Thus, if the guidelines for cell number and cell density are followed for Tier 1 and Tier 2 lines, analysis by FACs to determine cell cycle state is not required.
- **Presence of mycoplasma.** Cell cultures should be tested periodically for the presence of mycoplasma. This is particularly critical if the growth of cells is altered. The ChIP, DNase, FAIRE, DNAme standards July 2011 mycoplasma testing protocol used by Bionique, which does mycoplasma testing for ATCC, is recommended.

● **Freezing cell aliquots.** Each ENCODE group should freeze a viable aliquot of each cell type used for any experiment for potential future phenotyping. The cells should be stored in the laboratory in which they are frozen.

II) ENCODE Standards for eCLIP-seq Experiments

There is considerable diversity in the way CLIP-seq experiments are designed, executed, scored, and reported, with numerous protocol variants and experimental designs described in the literature. There are also substantial differences in the number of sites detected for each factor, ranging from hundreds to hundreds of thousands. This is partly a function of the underlying biology of different factors, but it is also a function of differences in the quality, scoring and reporting of experiments. To address the needs for reproducible high quality data and to facilitate analysis and dissemination of results, we have worked to develop standards and best practices for CLIP-seq experiments. We have based this framework on the recently described eCLIP-seq method (Van Nostrand *et al*. 2016), as it enables robust profiling of RBP targets and incorporates a size-matched input for proper input normalization and removal of common artifacts. Informed by our results profiling over one hundred RBPs, here we describe our recommendations on study design, quality control, evaluation of results, reporting and archiving. Below are our current standards with the expectation that they will be revised as protocols and technologies change over time.

IIa. Antibody characterization

A key to any CLIP-seq experiment is the identification of antibodies that can successfully immunoprecipitate the targeted RBP in CLIP-seq conditions. CLIP-seq experiments have been performed successfully using both polyclonal and monoclonal antibodies. However, the success of these experiments is heavily dependent upon the quality of the antibodies, which can vary considerably in terms of specificity and performance. Consequently, we propose a set of standards for antibody characterization. The thresholds used in these standards, while somewhat arbitrary, provide a useful guide for helping to ensure that high quality data are generated.

The workflow for characterization of a new antibody that binds an RBP is summarized in Figure 1; both a primary and secondary characterization are required for each new antibody and for a new lot of a previously characterized antibody (Table 1).

IIa.1. Primary mode of characterization

Antibodies are characterized by one primary method:

1. Immunoblot analyses: Immunoblot analysis should be performed on protein lysates from whole cell extracts and immunoprecipitated material. The primary reactive band should contain at least 50% of the signal observed on the blot. Ideally, this band should correspond to the size expected for the protein of interest. However, the mobility of many factors deviates significantly from the predicted size due to modifications, isoform differences or intrinsic properties of the factor. Therefore, main bands that differ from the expected size by more than 20% or multiple bands are acceptable under certain circumstances, such as if the unexpected mobility has been properly documented in published studies using the same antibody lot, if the signal in the band is reduced by shRNA knockdown or genetic knockout, or if the factor can be identified in this

band by mass spectrometry. Before proceeding to CLIP assays, it is helpful to demonstrate that the protein of interest can be efficiently immunoprecipitated.

IIa.2. Secondary mode of characterization

In addition to the primary mode of characterization, a secondary method is performed using the following assay:

1. Knockdown or knockout of the target protein. Immunoblots are performed using extracts from shRNA knockdowns or from knockout mutants. The primary immunoblot signal, along with additional immunoreactive bands, should be reduced to no more than 50% of the original signal.

IIa.3. Other considerations

1. Additional controls are welcomed but cannot replace the criteria indicated above. These include competition of immunoblot, immunofluorescence and/or CLIP signals using peptide and/or protein for the RBP of interest. Signals are expected to be diminished. Correlation of the pattern of protein expression at different developmental timepoints or different conditions using immunoblot analysis or immunofluorescence with that expected for the protein based on RNA expression is also a useful criterion.
2. For antibodies directed against members of a multi-gene family, antibodies should be prepared to protein regions that are unique to individual family members. Any potential cross-reaction should be noted when reporting data collected using that antibody.
3. For antibodies that have been previously characterized for one cell type, only one validation method is required when used for CLIP-seq with a new cell type or organism. If an antibody has been validated in at least 3 different cell types, no further validation is needed for use with additional cell types.
4. Characterized antibodies may be used by different groups without further characterization, assuming they are derived from the same lot. Antibodies from different lots must be characterized as if they were new antibodies.

IIb. CLIP-seq Data Production Standards

In order to ensure that experiments are reproducible and high quality, standards have been established for performance of eCLIP-seq experiments.

Sequencing depth

For eCLIP-seq experiments, although the number of targets identified varies substantially depending on the factor, antibody, and the algorithm used for peak calling, it also depends on the depth to which the sample is sequenced. For practical purposes and cost considerations for mammalian cells, we defined a minimum depth of 1 million uniquely mapped, non-PCR duplicate (hereafter referred to as “usable”) reads per replicate as a standard for RNA binding proteins. This standard was higher than most published CLIP-seq datasets, and in our experience yields sufficient depth to perform peak calling for RBPs with varying binding profiles. Datasets with lower read depth can be rescued if it is shown that the factor binds to a small number of regions that has already saturated at the 1 million usable read depth, or that the uniquely mapped read number is low due to substantial binding to multi-copy regions (for example, rRNA, snRNA, or repetitive elements including retrotransposons). It is recommended that the number of reads obtained for each replicate be similar. Input samples (see below) need to be sequenced to a similar or greater depth.

Size-matched input controls

A paired size-matched input library was generated as a control experiment for each eCLIP-seq experiment. There exist many potential sources for bias in CLIP-seq signal, including sequence preferences in UV crosslinking, RNase fragmentation, adapter ligation, PCR amplification, and read mappability (sequence uniqueness in the genome), which are best controlled by a paired input library. Furthermore, as the CLIP-seq procedure includes size-selection after transfer to nitrocellulose membranes (selecting a region spanning ~75 kDa above the bound RBP, which corresponds to RNA fragments of 200-250 nt in length), the input should be size-matched to this region to ensure proper normalization and comparison. Controls should be performed per cell batch, as underlying changes in gene expression can significantly alter observed binding. For practical purposes and cost considerations, for each RBP assayed as part of the ENCODE project, one size-matched input experiment was performed, where the input sample was taken from one of the two RBP biosamples pre-immunoprecipitation. Individual users are advised to perform paired size-matched input experiments for each biosample if feasible.

Peak identification and scoring

Scoring procedures for eCLIP data are still being actively developed. Our current standard algorithm for calling and scoring peaks performs initial identification of read clusters using CLIPper (Lovci *et al.*, 2012) followed by comparing each of these clusters against the size-matched input to determine significantly enriched peaks. If other peak-calling methods are used, it is critical to incorporate the paired input in order to identify regions of significant enrichment. We generally recommend that for a high stringency set of high-confidence peaks, cutoffs should be set at fold-enrichment ≥ 8 and p-value ≤ 10^-3; however, these can be relaxed for various analyses, as there remain true positive peaks below these thresholds. Visual inspection of known targets can also be used to determine proper cutoffs for individual datasets.

Reproducibility across biological replicates

Reproducibility is assayed in two ways. First, at a broad level, there should be high concordance between replicates for the fraction of peaks identified in introns / 3'UTR / coding sequence / 5'UTR, and (if present) relative binding to multicopy or repetitive elements (e.g. rRNA, tRNA, SINE/LINE, etc.).

At the peak level, reproducibility is assayed by standard Irreproducibility Rate (IDR) analysis (Li *et al.* 2011). We have found that sorting peaks by fold-enrichment in CLIP relative to size-matched input works well for IDR analysis (Van Nostrand *et al.*, 2016). To automatically pass quality-control, replicates should have a rescue ratio and self-consistency ratio of less than two. However, we have found that these metrics are inadequate for some RBPs that bind repetitive elements, and that these can be accepted as specific exceptions if they do not meet these criteria.

There are a wide variety of RBP binding types, ranging from RBPs that bind to a narrow region centered on a specific sequence motif to those that coat entire transcripts or transcript regions. However, we note that we have found that the above methods can be used for both broad and

narrow binders, as shearing, ligation, and amplification biases typically lead to reproducible narrow 'peak' calls even in broadly bound regions. However, for extremely broad binders (e.g., RNA pol II or ribosomal components), reproducibility can also be assayed at the region- or transcript-level.

Number of replicates

In order to show that experiments are reproducible, at least two biosamples must be analyzed. Because peak identification is somewhat dependent on read depth, it is recommended, but not required, that the number of mapped reads be within two fold of each other.

Other metrics

Although not required for submission, there are other quality control metrics suggested to be performed on eCLIP data before submission: 1) Visual inspection of genome browser tracks to assess data quality and concordance between replicates, 2) Analysis of the binding sites within transcript regions (e.g. intronic versus exonic), or specific RNA targets (e.g. rRNA or tRNA), relative to known roles of and sub-cellular localization of the RBP, and 3) Comparison of sequence motifs identified *de novo* within peak regions relative to previously known binding motifs. There are many reasons why these metrics may not work for individual RBPs, but they can provide additional confidence to datasets that conform to previous knowledge of the RBP.

III) Recommended standards for reporting eCLIP-seq data

In order to facilitate data sharing among laboratories both within and outside the consortium, and to ensure that the results can be reproduced, requirements have been established for data sharing.

Storing high throughput sequencing data

- Image files from sequencing experiments do not need to be stored.

Submitting ENCODE eCLIP-seq data

- Raw data (read1 and read2 .fastq files) should be submitted to both GEO and the DCC for biological replicates 1 and 2 and size-matched input. Data should be flagged as part of the mod/ENCODE project through the use of the appropriate genome project ID

- Processed data should be submitted to the relevant DCC as:
  - "Usable read" (uniquely mapped, post-PCR duplicate removal) mapping information (.bam file) for biological replicates 1 and 2 and size-matched input
  - Signal tracks (bigwig format) for biological replicates 1 and 2 and size-matched input
  - Called, input-normalized peaks (see below) for biological replicates 1 and 2
  - Metadata, including peak caller version used (see below)

Target Region and Peak Calling for eCLIP-seq

Common features that should be reported to the DCC are:

- Cluster, defined as a region identified by the CLIPper algorithm as enriched above transcript-level background
- Position (chromosome, strand, and peak start and end position)
- Significance statistics:
  - 1: Signal value (e.g., log2(fold-enrichment)) for RBP IP compared to size-matched input for the listed peak region
  - 2: P -value (typically reported as –log10(p-value)), determined by Fisher's Exact test (or Chi-Square approximation where appropriate) comparing RBP IP versus size-matched input for the listed peak region

IV) Metadata

The DCC requires submission of basic experimental data, including minimally the following:

1. Investigator, organism or cell line, experimental protocol (or reference to a known protocol).
2. Indication as to whether an experiment is a technical or biological replicate (at least two must be performed).
3. Catalog and lot number for any antibody used. If not a commercial antibody, precise source and method of preparation of the antibody.
4. Information used to characterize the antibody, including summary of results (images of immunoblots, etc.) and link to antibody validation information submitted to DCC.
5. Peak calling algorithm and parameters used, including thresholds and reference genome.
6. Criteria that were used to validate the quality of the resultant eCLIP-seq data, including any exceptions appropriate for the RBP.
7. If the experiments fails to meet any of the standards an explanation is required.

The ENCODE metadata listing the peak callers (and versions) that have been used will in turn link to the websites maintained by the individual groups that allow for the downloading of peak caller software by outside data users. Data producers are expected to update information about peak calling software (including versions) on their websites as soon as new or updated software is implemented.