ENCODE Guidelines for Experiments Generating ChIP-seq Data
January 2017

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 DNA-associated proteins.

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.

- **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 x 10^5 cells/ml and 1 x 10^6 cells/ml.
    - K562 cells should be grown to a maximum density of 7-8 x 10^5 cells/ml.
    - HepG2, MCF-7, A549, HEK293, and SK-N-SH cells should be grown to a maximum of 70-80% confluence.

- **Cell cycle and gene expression state.** Different laboratories performed gene expression experiments 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.

- **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 ChIP-seq Experiments

Despite their widespread use, there is considerable diversity in the way global ChIP experiments are designed, executed, scored, and reported. There are substantial differences in the number of sites detected for each factor, ranging from hundreds to tens of thousands. This is partly a function of the underlying biology of different factors and modifications, 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, the ENCODE consortia have worked to develop standards and best practices for ChIP-seq experiments. Because the methods for performing genome-wide ChIP continue to evolve, our standards and practices have also evolved during the course of these projects, informed by results from over one thousand experiments performed in different organisms (e.g. *D. melanogaster*, *C. elegans*, mouse, and human), and multiple independent data production pipelines. The resulting guidelines include recommendations on study design, quality control, evaluation of results, reporting and archiving.

IIa. Antibody Characterization and Epitope Tagging

ChIP 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, which differ for antibodies generated against transcription factors relative to those used to study histone modifications. The thresholds used in these standards, while somewhat arbitrary, provide a useful guide for helping to ensure that high quality data are generated. Most recent antibody characterization guidelines can be found in the ENCODE portal (https://www.encodeproject.org/about/experiment-guidelines/#antibody).

Characterization of antibodies directed against transcription factors

For validation of an antibody that binds a transcription factor both a primary and secondary characterization are required for each lot. For more detailed information, please visit the newest Antibody Characterization Standards for transcription factors found in the ENCODE portal (https://www.encodeproject.org/documents/c7cb0632-7e5f-455e-9119-46a54f160711/@@download/attachment/ENCODE_Approved_May_2016_TF_Antibody%20Characterization_Guidelines.pdf).

Epitope-tagged proteins

The creation of cell lines or organisms expressing an epitope-tagged transcription factor for ChIP is a useful alternative and is particularly valuable in cases where a ChIP quality antibody is not available against the endogenous protein. Please see the document for Epitope-tagged transcription factor ChIP-seq on the ENCODE portal for more

information on how to validate cell lines containing epitope-tagged proteins (https://www.encodeproject.org/documents/35a9f776-dd6a-44e3-8795-50ead83f34f7/@@download/attachment/Guidelines_for_Use_of_Epitope_Tags_in_ChIP-seq_Jan_2017.pdf).

Histone Modifications

All commercial histone antibodies must be validated by at least two independent methods and new lots of antibody must be analyzed independently. For more detailed information, please peruse through the Antibody Characterization Standards for histone modifications and chromatin-associated proteins (https://www.encodeproject.org/documents/4bb40778-387a-47c4-ab24-cebe64ead5ae/@@download/attachment/ENCODE_Approved_Oct_2016_Histone_and_Chromatin_associated_Proteins_Antibody_Characterization_Guidelines.pdf).

IIb. ChIP-seq Data Production Standards

In order to ensure that experiments are reproducible and high quality, standards have been established for performance of ChIP-seq experiments (https://www.encodeproject.org/data-standards/chip-seq/). The ENCODE Consortium has provided specific definitions for these terms, which could be found at the portal (https://www.encodeproject.org/data-standards/terms).

Sequencing Depth

For ChIP-seq experiments, 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 current read depth requirements please visit https://www.encodeproject.org/data-standards/chip-seq/.

Controls

Control experiments must be performed. Breakage during sonication can occur preferentially in regions of open chromatin resulting in non-uniform background signal. In addition, many cell lines are aneuploid and have many large regions of genomic duplications, which can heavily influence peak sizes and rankings. Control DNAs include “Input” DNA, in which DNA is isolated from cells that have been crosslinked and sonicated under conditions similar to the experimental sample. In cases where factor binding to DNA is environmentally induced, (e.g. after hormone induction of the glucocorticoid receptor or when analyzing a protein expressed from an inducible promoter), a potential appropriate control is a parallel ChIP experiment done on cells in the uninduced conditions. For epitope-tagged constructs, wild-type cells lacking the epitope tag or the input DNA of the tagged cell line can be used as controls. If amplification is used to prepare the experimental sample, then the control DNA must be prepared using the same amplification procedure; note that biases in amplification can increase the chances of overrepresentation and underrepresentation of sequences. Control experiments need to be performed for each cell line, developmental stage and

different condition/treatment since the open chromatin regions are likely to change under the different cell types, stages, and conditions.

Number of Replicates

In order to ensure that experiments are reproducible at least two replicates must be performed. For ChIP-seq experiments with RNA Pol II, the use of more than two replicates did not significantly increase the number of targets identified. Therefore, it is recommended that ChIP-seq experiments be performed in duplicate using two independent samples (i.e. biological replicates). Samples with limited availability of experimental material might be exempted.