

BETA CELL BIOLOGY CONSORTIUM

ChIP-Seq Minimal Standards and Guidelines



Version 1 - Last modified on May 3, 2011



This document is derived from ENCODE and modENCODE standards and is provided by the BCBC Bioinformatics and Epigenomics Workgroup. Please contact the workgroup with suggestions or questions.

CONTENTS

A. Background.....	1
B. Antibody characterization standards	1
C. Antibody characterization standards for antibodies against histone modifications	2
D. Epitope tagging.....	2
E. ChIP-Seq experimental parameters.....	3
F. Specifics of data to be reported	4
G. Figures	5

A. BACKGROUND

Chromatin Immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) have become valuable approaches for the global mapping of transcription factor binding sites and histone modifications (Figure 1). Despite their widespread use, considerable differences exist in how the experiments are conducted, how the results are scored, and how the data are archived for public dissemination. The ENCODE and modENCODE Consortia have developed working standards for the performance and distribution of results of ChIP-seq experiments. Summarized below are relevant standards that have been adapted for use by the beta cell consortium, including guidelines for antibody validation, numbers of experimental replicates, and sequencing depth as well as parameters that should be reported with each experiment. It is recognized that these parameters may change over time both as technology and our understanding improves.

B. ANTIBODY CHARACTERIZATION STANDARDS

All antibodies must be characterized by one primary assay and one secondary assay and the results reported before any data using that antibody can be deposited (Figure 2).

Primary characterization

1. *Immunoblot analysis of total cell, nuclear, or chromatin extracts* - at least 50% of signal must reside in a single band of expected size, **OR** if that fails,
2. *Immunofluorescence analysis* - nuclear signal is present in cells expressing the factor.

Secondary characterization

1. *shRNA/siRNA knockdown or knockout strain* - signal for immunoreactive bands or immunofluorescence must diminish by >70%, or for known targets measured by ChIP must diminish by >50%, **OR**
2. *Immunoprecipitation/Mass spectrometry* - the primary immunoreactive band must contain the factor of interest and chromatin-associated proteins not known to form a complex with the factor cannot be present in immunoreactive bands, **OR**
3. *ChIP with multiple antibodies against different parts of the same protein or members of the same complex* - agreement between experiments must meet the standard used for comparing biological replicates (see below), **OR**
4. *ChIP from cells expressing epitope-tagged factor* - agreement between experiments must meet the standard used for comparing biological replicates (see below).
5. *Motif analysis*. – if a factor has a well characterized motif from in vitro binding studies or another justifiable criterion, and either no paralogs expressed in the cell lines being studied or the antibody is raised to a unique part of the protein then motif enrichment can be used. In this case the motif should be in 70% or more of the binding regions and enriched over the genome with a p value of <10exp5.

C. ANTIBODY CHARACTERIZATION STANDARDS FOR ANTIBODIES AGAINST HISTONE MODIFICATIONS

A specific set of characterization requirements exist for this class of antibody.

Primary characterization (required)

Immunoblot analysis on nuclear extract and recombinant histone- expected band must constitute >50% of signal and at least 10-fold enrichment over any other band in nuclear extract, expect at least 10-fold enrichment for recombinant modified histone over unmodified protein.

Secondary characterization (choose one)

1. *Peptide binding test* - binding to peptide with modification of interest must show at least 10-fold enrichment relative to peptides with other modification, **OR**
2. *Immunoprecipitation/Mass spectrometry* - target modification must constitute >80% of signal **OR**
3. *Mutant strains defective in modified histone* - immunoblot signal should be reduced by >90% in mutant compared to wild-type, **OR**
4. *Histone with mutated site of modification* - at least 10-fold reduction of immunoblot or ChIP signal from mutant protein as compared to wild-type.

D. EPITOPE TAGGING

Cells lines should be constructed that contain tagged proteins expressed near endogenous levels. This can be accomplished using with proteins expressed from their own promoter or from other promoters; for the latter immunoblot analysis should be performed to demonstrate that proteins are near endogenous levels. If protein cannot be detected, overexpression can be used but must be indicated. Ideally one should show that a tagged protein is still functional but we recognize that this is not always easy to carry out. Control

experiments for epitope tagged lines include performing parallel ChIP-Seq experiments from untagged strains.

E. CHIP-SEQ EXPERIMENTAL PARAMETERS

1. Read depth and sequencing quality

- a. At least 10 million uniquely mapped reads/replicate are required for mammalian samples, and at least 2 million for worms or flies. For broad chromatin marks (e.g. H3K27me3) 20 million uniquely mapped reads/replicate are recommended.
- b. The number of reads for biological replicates should be within a factor of 2.

2. Controls

- a. Either input DNA (sonicated, reverse crosslinked DNA that has not been immunoprecipitated) or IgG (DNA obtained by immunoprecipitation with a non-specific antibody fraction) may be used as a scoring control;
- b. Control DNA must be amplified identically to experimental material;
- c. A separate control is required for each cell type, cell treatment, or developmental stage being studied
- d. Controls must be sequenced to same read depth as experimental samples.

3. Number of replicates

- a. Two biological replicates are required for each dataset;
- b. For each replicate, targets are identified (see below) and 80% of the top 40% of peaks that exceed a given threshold (usually a false discovery rate of .01) must be present in the list of identified peaks for the other replicate;
- c. The number of targets identified for each replicate cannot differ by more than a factor of 2;
- d. Reads from replicates which meet these criteria are usually combined and the data rescored.

4. Scoring

Any scoring algorithm that accounts for control signals is acceptable for peak-calling. PeakSeq and MACS are commonly used by ENCODE.

5. Data that do not meet the criteria

If, after repeated attempts, data does not meet the criteria above, data may be released with a note indicating that the criteria have not been met and explaining why the data is released without meeting criteria.

6. Data reporting

The following must be reported (Detailed description below):

- a. General information including a contact person, objectives of the experiment and experimental factors.

- b. Descriptions of the samples (cell lines and conditions), which are biological replicates, and relation to sequence runs performed.
- c. How the antibodies were characterized.
- d. The methods used to perform the experiments, those generating the sequence, and those used for scoring. Ideally entire enrichment lists are reported; however, minimally lists with an FDR less than 1% must be presented.
- e. Sequencing reads for each experiment and control along with scored list of targets must be submitted to a public archive (such as NCBI GEO or ArrayExpress).
- f. For each scored target the following parameters must be reported: signal strength (either total peak figure or maximum peak signal), significance, genomic location and genome build used for mapping reads and coordinates. Also whether the scored target was replicated.
- g. Exceptions to the rules indicated above can be made but must be flagged.

7. Broad peak data

It should be noted that for factors that give broad peaks (Pol2, H3K27trimethylation), unified methods for scoring and assessment of reproducibility have not been established. Investigators should report the criteria that they use.

F. SPECIFICS OF DATA TO BE REPORTED

1. Experiment

- a. Contact person,
- b. Objective,
- c. Experimental factors.

2. Samples

- a. Description: such as (where applicable):
 - i. Cell line, Lot number
 - ii. Cell or tissue
 - iii. Mouse strain, relevant alleles
 - iv. Gender, age, disease status
- b. Protocols used,
- c. Identify biological vs. technical replicates.

3. Antibody Characterization

- a. Company/ Core, catalog or ID and lot number
- b. Describe methods used and quality assessment; provide images.

4. Sequence

- a. Protocols used (a protocol should be made available on the BCBC3.0 web site),
- b. Relation to samples,
- c. Machine (version),
- d. Sequence length,

- e. Single or paired,
- f. Data file of sequence reads with quality scores (e.g. FASTQ).

5. Sequence Analysis

1. Quality assessment (correlations between biological replicates),
2. Reference genome,
3. Mapping software (version) and parameters used,
4. Numbers of reads, whether trimming was done: total, mapped, unique,
5. Methods used to call peaks,
6. Peak coordinates,
7. Peak signal,
8. Confidence score (e.g., p-value).

G. FIGURES

Figure 1. ChIP-seq/ChIP-chip workflow. Steps for which specific standards are presented in this document are indicated in red. For other steps, standard ENCODE protocols exist which should be validated and optimized for each new cell line/tissue type or sonicator. * indicates a commonly used but optional step.

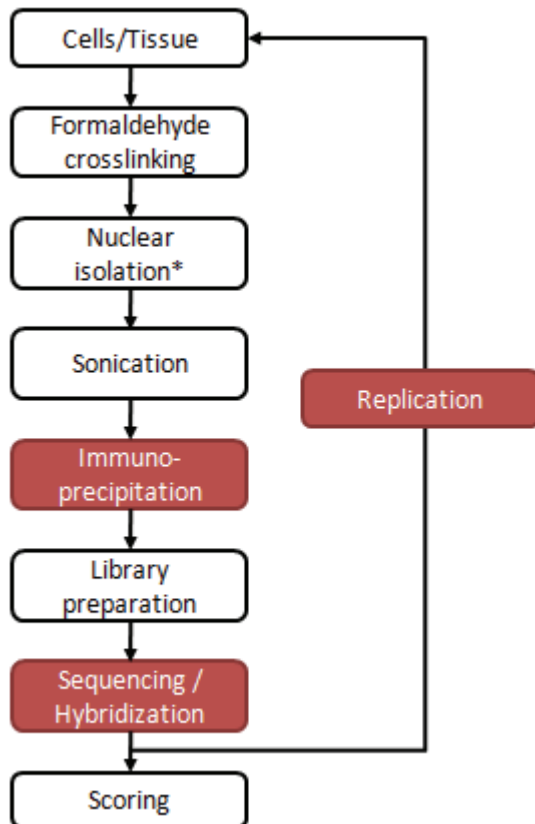


Figure 2. Flowchart for antibody characterization

