Recommended Standards for Reporting mod/ENCODE Data
Version 1.0

January 23, 2009

Storing High Throughput Sequencing Data

- Image files from sequencing experiments do not need to be stored for the long term.
- Sequencing intensity files do not take up a lot of storage space and should be saved for the long term.

Submitting mod/ENCODE Data

For ChIP-chip, DNase-chip/array, and FAIRE-chip:

- Raw data should be submitted to GEO.
    - Data should be flagged as being part of the mod/ENCODE project upon submission to NCBI
- Processed data should be submitted to the relevant DCC as:
    - Ratio tracks
    - Called peaks (see below)
    - Metadata, including peak caller version used (see below)

For ChIP-seq, DNase-seq and FAIRE-seq:

- Raw data should be submitted to GEO (1)
    - Data should be flagged as being part of the mod/ENCODE project through the use of the appropriate genome project ID
    - Each replicate should be submitted independently
- Processed data should be submitted to the relevant DCC as:
    - Input signal or alignments (2) (will not be done for DNase/FAIRE Tier 3 lines)
    - ChIP, DNase, and FAIRE signal or alignments (2)
    - Interpreted data signal
    - Called peaks (see below)
    - FASTQ files (3)
    - Metadata, including peak caller version used (see below)
- The DAC will propose how to generate gene target lists for ENCODE data. These lists will be submitted to the DCC in the future.

Target Region and Peak Calling for ChIP, DNase and FAIRE Experiments

Point Source Peaks

For point source peaks (e.g., DNase, FAIRE, or signals from ChIP experiments with antibodies to sequence-specific transcription factors), common features that should be reported to the DCC are:

- Peak, defined as a single base pair
- Start and end, defined as specific base pairs
- Significance statistics using a three slot model (the inclusion of slots 2 and 3 is optional for data submitters):
- Slot 1: Signal value (e.g., fold enrichment) using an algorithm chosen by the submitter
- Slot 2: P -value determined using a method chosen by the submitter
    - Slot 3: Q-value (false discovery rate correction) determined using a method chosen by the submitter
- Metadata, including peak caller approach used (see below) and methods for determining signal values, P-values, and Q-values, as applicable

Broad Regions (does not pertain to DNase and FAIRE)

- Start and end, defined as specific base pairs
- Significance statistics using a three slot model (the inclusion of slots 2 and 3 is optional for data submitters):
- Slot 1: Average signal value across region (e.g.., fold enrichment) using an algorithm chosen by the submitter
- Slot 2: P -value determined using a method chosen by the submitter
    - Slot 3: Q-value (false discovery rate correction) determined using a method chosen by the submitter
- Metadata, including peak caller approach used (see below) and methods for determining signal values, P-values, and Q-values, as applicable
- Point-source peaks can be called in addition to broad regions (i.e. one can have "peaks" and potentially "valleys" within "regions").

It is up to the investigator to determine whether their data best fits the broad region/point source peak data or both.

Metadata for Peak Caller

As an overall aim, mod/ENCODE should strive to provide public access to peak calling software so outside data users can replicate the findings of the mod/ENCODE data producers. Currently, peak calling software can be downloaded from the websites of the individual data production groups and initially the consortium and should build on this arrangement. In the longer term, it may be worthwhile to attempt to standardize peak and region calling approaches so as to achieve optimal integration of data.

The metadata for each experiment should include a free form field where data producers are required to include information about the peak caller version that was used to produce the hit list along with any information on parameters used for a particular experiment.

The modENCODE metadata field and ENCODE track documentation will be linked to a DCC web page that lists the peak callers (and versions) that have been used for mod/ENCODE data, which 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.

Notes:

(1) NCBI intends for sequence data to be submitted directly to GEO. GEO will pass primary sequence data to the SRA. If more convenient, data producers can make dual submissions to the SRA and GEO. If this is done, sequences should be submitted to the SRA and metadata, processed data, and a link to the SRA accession should be submitted to GEO.

(2) The ENCODE DCC will accept sequence alignments and/or signal graphs for input and ChIP data. Sequence alignments will be posted for download, and will be loaded and displayed as a 'Counts graph' (count of tags overlapping each base) if no signal graph is submitted.

(3) The ENCODE DCC will host FASTQ format sequence data and, where feasible, provide co-located cluster computing for analysis since the GEO/SRA pipeline and access tools are not mature enough at this stage to adequately support ENCODE analyses.