DNase-seq Data Standards
January 2017

I. Biological Replicate Samples

Experiments should have at least two independently derived biological replicates of the same cell/tissue type/state.

- Replicates may be isogenic (in the case of cell lines or primary cell samples from the same individual) or anisogenic (same cell/tissue type/state but from a different donor).
- In cases of anisogenic datasets – i.e., where strictly matched biological replicate samples are not obtainable from primary samples due to limited availability, or where high-quality data cannot be generated due to primary sample quality – the requirement for a second biological replicate may be waived

II. Primary Sample Processing Criteria/Restrictions

- The read length should be a minimum of 36 base pairs.
- Read trimming of adapter sequences is recommended.
    - Failure to trim sequences of fragments sized shorter than the sequencing cycle number can lead to filtering small fragment signal and create bias in the resulting alignments.
    - There is no post trimming read length minimum, but effective fragment mapping drops significantly at kmer-lengths of less than 22 base pairs.
- Adapter sequences used in library creation should be documented and available to the pipeline.
- Barcodes/UMI coding should be indicated in the metadata and available for accurate application of duplication filtering methods by the pipeline.
- Sequencing may be paired- or single-end, as long as sequencing type is specified and read pairs are indicated. Paired-end sequencing is preferred.
- The sequencing platform used must be indicated.
- Alignment files are mapped to GRCh38 in case of human samples, or mm10 in case of mouse samples.

III. Data quality metrics

Data quality metrics should report (1) the proportion of duplicate reads; (2) the proportion of reads deriving from the mitochondrial genome; and (3) the proportion of uniquely mapping reads exclusive of mitochondrial reads.

- Libraries should contain no more than 5% duplicate reads
- Libraries should contain no more than 10% mitochondrial reads
- Libraries should optimally yield >80% uniquely mapping reads, but in no case less than 75% uniquely mapping reads (both excluding mitochondrial reads)

IV. Quantification of Data Quality

A Signal Portion of Tags (SPOT) score must be calculated and provided as primary metadata for all samples.

- SPOT scores should be computed from at least 20 million uniquely mapping reads
- SPOT scores should be calculated on de-duplicated data or on data for which the

duplicate rates are less than 5%.

- SPOT scores of 0.4 or higher define high-quality reference data
- SPOT scores of 0.3-0.4 represent intermediate quality data and are acceptable when primary material is limiting, and/or better quality data cannot be obtained from a primary tissue.
- SPOT scores of 0.25-3 represent lower quality data and are acceptable when primary material is limiting, and/or better quality data cannot be obtained from a primary tissue.
- SPOT scores < 0.25 represent unacceptably low data quality that can significantly confound analyses of individual data sets. Such data may be made available, with appropriate cautions, only in the case of the most difficult to procure primary tissues for which no other data are available.
- Any sample with SPOT score of <0.3 should be targeted for replacement with a higher quality sample, whenever possible
- As DNase-seq produces stereotypical fragments, general measures of complexity are less informative than they are in other assays where random shearing is more prevalent. Please see the DNase-seq data standards page for acceptable levels of duplication for different library preparation protocols.

V. Target sequencing depth

- The target depth for comprehensive delineation of DNase I hypersensitive sites (DHSs) in high quality DNase-seq samples is ~50 million uniquely mapping reads.
- The minimum acceptable depth for delineation of DHSs is 20 million uniquely mapping reads.
- The target depth for comprehensive delineation of DNase I hypersensitive sites (DHSs) in high quality DNase-seq samples is 150-200 million uniquely mapping reads.
- The minimum acceptable depth for delineation of DNase I footprints is 100 million uniquely mapping reads
- Samples with SPOT scores of <0.3 are generally not suitable for comprehensive detection of DNase I footprints, regardless of sequencing depth.

VI. Replicate Concordance

- Samples should have a gene-level Pearson correlation of >0.9 between isogenic replicates and >0.85 between anisogenic replicates.

VI. Metadata

Experiments must pass routine metadata audits to be released.