Standards and Guidelines for Whole Genome Shotgun Bisulfite Sequencing

I. Introduction.

Shotgun sequencing of genomic DNA subjected to sodium bisulfite conversion (MethylC-Seq) has enabled single-base resolution, strand specific identification of methylated cytosines throughout the majority of the genome of several eukaryotic organisms. With recent increases in high-throughput sequence yield, routine generation of high-coverage whole-genome mammalian DNA methylomes is now feasible. This document aims to outline standards in experimental methodology, sample and experimental recording, and data analysis that will guide the production of high quality DNA methylomes via shotgun bisulfite sequencing. Due to rapid methodological and technological advances, this document should be revised annually.

II. Information to be supplied with each sample used for a MethylC-Seq experiment.

MethylC-Seq data should be accompanied by information concerning the biological source of the genomic DNA (gDNA) and the protocols used to extract the gDNA.

1. For cell lines the following information should be recorded and provided:
    a. Cell line source and lot number.
    b. Growth time/passage number.
    c. Cell density.
    d. Protocol used to culture cell lines.
    e. Results of tissue culture contaminant (e.g. mycoplasma/wolbacia) tests, if conducted.
    f. Results of karyotyping and marker analysis, if conducted.
    g. Confirmation of freezing cell aliquots of examined lines.
2. For sub-cellular compartments, tissues, organs or whole organisms, the following should be recorded and provided:
    a. Protocols for synchronization of animals, for purification of tissue or cell types.
    b. Amounts of starting material (tissue/organ weights, cell number, etc).
3. Protocols used to isolate gDNA.
4. Fragmentation, end-repair protocol (if used) and experimentally-based estimation of library generation input gDNA fragment size distribution.
5. Quantitation of gDNA during sequencing library production. Methodologies and measures should be recorded for the amount DNA input into the library preparation and the bisulfite conversion procedure.
6. Quantity of unmethylated Lambda gDNA or alternative control sequence spiked into sample prior to gDNA fragmentation (% w/w).

III. Performance of MethylC-Seq Sequence Experiment: Number of replicates and sequencing depth.

1. Replication: In order to ensure that the data are reproducible, experiments should be performed with two or more biological replicates, unless there is a compelling reason indicating that this is impractical or wasteful (e.g. overlapping time points with high temporal resolution). A biological replicate is defined as an independent growth of cells/tissue and subsequent analysis. Technical replicates of the same library are not required but may be useful to reduce unnecessary post-sequencing removal of sequence reads displaying coincident alignment positions that may be indicative of potential PCR clones.
2. Sequencing depth. A full DNA methylome should have at least 30X coverage of the genome when reads from biological replicates are combined. For example, a methylome with 2 biological replicates, each with 15X coverage, would be sufficient. Due to strand specificity of bisulfite sequencing data, 30X coverage is equivalent to 15X per strand of the genome. In addition to genome coverage, the average coverage of CpGs may be a useful measure for sequencing depth.

IV. Information supplied concerning steps taken prior to the sequencing reactions.

1. *Method of fragmentation of gDNA Samples for Sequencing*. The method of fragmenting gDNA prior to sequencing library construction must be described in the metadata supplied with the sequencing results. Either single-end or paired-end sequencing approaches are suitable, but provision of average input gDNA fragment length is informative for sequence analysis considerations.
2. *Confirmation of unmethylated Lambda gDNA or alternative spike-in prior to library construction*. Unmethylated *c*I857 *Sam7* Lambda DNA (Promega Cat# D1521) must be spiked into the sample gDNA prior to fragmentation. Typical spike-in levels range from 0.1 - 0.5% (w/w), and this quantity should be recorded.
3. *Methylated adaptor characteristics*. Source of methylated adapters that were ligated to the fragmented gDNA (custom-made/3rd party-made), their sequence and any other modifications.
4. *Method of bisulfite conversion*. The method of bisulfite conversion, including reagent/kit supplier(s) and procedure.
5. *Other information*. This information should detail whether the sample consists of pooled and bar coded MethylC-Seq libraries, the estimated depth of sequencing (e.g. number of reads passing quality filtering prior to alignment), lengths of the reads, whether the reads are intended to be single

or paired end, description of the control lane sample used for estimation of matrix and phasing (for Illumina sequencing, e.g. phiX174 control).

V. Information supplied concerning pre- and post-sequencing mapping, read statistics and quality scores.

1. *Pre-mapping data filtering/handling details*. Details must be provided of data analysis steps undertaken prior to read mapping. For example, trimming of low quality bases from reads, identification and removal of adapter sequences.
2. *Mapping of sequence data*. There are multiple short read mapping algorithms currently available that can natively handle bisulfite converted sequence alignment, or that can be used to align bisulfite converted sequence data through alternative approaches (e.g. C-free alignment). The use of multiple mapping algorithms requires that information regarding the mapping strategy and relevant mapping criteria are detailed.
3. *Mapping algorithm thresholds and settings*.
    a. Number of allowable mismatches, minimal score, maximum allowed sum quality scores at mismatches, etc.
    b. Were reads only allowed to match uniquely or were multiple genomic mapping positions allowed?
    c. For paired-reads, whether there are constraints regarding the pair mapping locations (within the same chromosome, within a certain genomic interval).
4. *Post-mapping data filtering/handling and results*.
    a. Clonal reads present in all reads derived from a single PCR reaction should be removed after mapping (single-end reads sharing the same 5' read alignment position or paired-end reads sharing both 5' read alignment positions).
    b. Any post-mapping filtering steps should be detailed (e.g. removal of inappropriately mapped reads from C-free alignments or 3' trimming of low quality mismatched bases).
    c. Paired-reads that have partial overlap in genome coverage should be trimmed from the 3' so as to avoid treating sequence derived from multiple passes of the same genomic DNA fragment as independent data points.
    d. Percent of raw reads mapped uniquely to the genome.
    e. Number of mapped reads and resulting genome coverage.
    f. Percent of genome and genomic cytosines covered.
5. *Empirical determination of bisulfite conversion frequency from lambda or alternative control as well as genomic sequence characteristsics.* Using the reads that map to the unmethylated lambda or alternative spike-in control it is necessary to report:
    a. Lambda genome or alternative spike-in coverage (% of bases covered and average depth).
    b. Percent bisulfite conversion at C, CG and non-CG (CH) sites within

the lambda genome.

c. Percent bisulfite conversion at CpC sites within the genomic DNA
d. Select a core set of promoter CpG islands (n=30 should be sufficient) as control regions for bisulfite conversion and determine the mCpA rate.