Assay Standards Working Group Recommendations, November 2012

Contents

Introduction

Recent technological advancements have enabled the reproducible assessment of epigenomic marks across the entire genome of human cells, and large-scale international efforts are now underway to generate high-resolution reference epigenome maps to accelerate the scientific exploitation of human epigenomic information. The epigenome maps thus generated integrate detailed DNA methylation, histone modification, nucleosome occupancy and coding and non-coding RNA expression in different normal and disease cell types, with the goal of providing new insights into many diseases, and the discovery of new means to control them.

The goal of the Assay Standards Working Group is twofold: to define the assays required for three distinct classes of reference epigenome, and to define standardized protocols and quality control (QC) metrics for each assay. The current members of the Working Group are as follows:

MARTIN HIRST, University of British Columbia/BC Cancer Agency, Vancouver, Canada;
IVO GUT, Centro Nacional de Analisis Genomico, Barcelona, Spain;
HIROSHI KIMURA, Osaka University, Osaka, Japan;
JOOST MARTENS, Radboud University, Nijmegen, The Netherlands;
KATSUHIKO SHIRAHIGE, University of Tokyo, Tokyo, Japan.

The guidelines described in this draft document are intended to provide a framework for the definition of reference epigenomes to be included within the International Human Epigenome Consortium. These recommendations are minimal standards based on current knowledge of the elements contributing to epigenomic regulation in human cells and the current state of epigenomic mapping technologies. In this regard it is expected that technological and research advancements will continue and thus these standards will need to be reviewed and updated annually.

Reference epigenomes submitted as part of IHEC must meet the minimum the criteria as listed in Table 1. Reference epigenomes generated as part of the NIH Human Reference Epigenome Mapping effort with be grandfathered as IHEC reference epigenomes as detailed in Table 1. All assays performed on reference epigenomes should conform to the guidelines listed below and the associated QC metadata submitted to the IHEC DCC in compliance with the standards developed by the IHEC Metadata Working Group.

The overarching goal of the International Human Epigenome Consortium is the generation of 1000 reference epigenomes. For this purpose a reference epigenome shall consist of data generated from a single human subject. Whenever possible a minimum of two biological replicates should be generated for a tissue or cellular target. The results of a pairwise comparison between the two replicates should serve as a guide for the generation of additional replicates. In cases where a single replicate is not possible, and a composite reference epigenome is generated from multiple subjects, the composite will be considered a single reference epigenome. A minimum of two composite reference epigenomes should be generated from non-overlapping datasets.

1: Reference Epigenome Criteria

Table 1. Reference epigenome definitions.

	Minimal IHEC Reference Epigenome	NIH Roadmap Minimal Reference Epigenome
Bisulfite-seq	REQUIRED	
MeDIP-seq (methylated DNA immunoprecipitation seq)		ANY OF 4 REQUIRED
MRE-seq (methylation-sensitive restriction enzyme seq)		
RRBS (reduced representation bisulphite seq)		
MethylCap-seq (methylation capturing seq)		
RNA-seq	REQUIRED	ANY OF 2 REQUIRED
Array based		
smRNA-seq (small RNA)	OPTIONAL	OPTIONAL
ChIP-seq input	REQUIRED	REQUIRED
H3K27me3	REQUIRED	REQUIRED
H3K36me3	REQUIRED	REQUIRED
H3K4me1	REQUIRED	REQUIRED
H3K4me3	REQUIRED	REQUIRED
H3K27ac	REQUIRED	OPTIONAL
H3K9me3	REQUIRED	REQUIRED
DNAseI hypersensitivity	ANY OF 3 OPTIONAL	ANY OF 3 OPTIONAL
DGF (Digital Genomic Footprinting)		
FAIRE-seq		

2: Whole Genome Shotgun Bisulfite Sequencing

2.1: Introduction

Shotgun sequencing of genomic DNA subjected to sodium bisulfite conversion (SBS-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.

2.2: Protocol: genomic DNA extraction

2.2.1: gDNA isolation

Standard gDNA quality measures should be employed. Currently the impact of gDNA fragmentation on the quality of resulting SBS-seq library is not known.

2.2.2: 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.

2.2.3: Control sequence spike-in

Unmethylated Lambda DNA (e.g. Promega Cat# D1521) or equivalent 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.

2.3: Protocol: MethylC-seq

2.3.1: 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 cells/tissue obtained from an independent human subject 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.

2.3.2: Sequencing depth

A full DNA methylome should have at least 30 fold redundant coverage of the reference genome from a single biological replicate. 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.

2.4: QC Metrics: MethylC-seq

2.4.1: 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.

2.4.2: 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.

2.5.3: 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).

2.5.4: 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. Evenness of coverage

g. Percentage of genome and genomic cytosines covered at >=1x, >=10x, >=30x.

h. Percentage of CpGs covered by more than ten reads.

i. For data originating from cancer genomes CNV status should be assessed by an orthologous methodology (array or whole genome shotgun) and the resulting data corrected to account for variance from diploid.

2.5.5: 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. Percent bisulfite conversion at C, CG and non-CG (CH) sites within the lambda genome.

b. Percent bisulfite conversion at CpC sites within the genomic DNA.

Using reads that map to the genome it is necessary to report:

c. Select a core set of promoter CpG islands (n=30 should be sufficient) as control regions for bisulfite conversion and determine the mCpA rate.

3: RNA-seq

3.1: Introduction

RNA-seq involves purification of RNA, followed by either selection of poly-A(+) RNA or depletion of ribosomal RNA. RNA is then converted into cDNA by one of two methods; 1) random priming, followed by cDNA fragmentation, end-repair and Illumina/SOLiD linker ligation or, 2) Enzymatic or chemical RNA fragmentation followed by linker ligation and cDNA generation. Following PCR amplification of tailed cDNA fragments with primers suitable for solid phase (Illumina) or emPCR (SOLiD) clonal amplification RNA-seq libraries are subjected to sequencing. Sequence alignment software is then used to compare the short sequence reads to reference genome and transcriptome databases, and exon-exon junction databases. From this analysis paradigm emerges data that is used for a variety of purposes, including the measurement of gene- level and exon-level expression abundance; detection of base changes (mutations and polymorphisms) relative to reference datasets; measurement of alternative splicing events; identification of gene fusion events; and identification of RNA editing events.

3.2: Protocol: total RNA extraction and QC guidelines

Some commercially available RNA extraction kits utilize columns that can deplete the small RNA fraction during total RNA extraction making the resulting RNA unsuitable for miRNA-seq library preparation. A standard Trizol based extraction methodology or a column run under non-selecting conditions (for example Ambion mirVANA) is recommended for REMC total RNA extraction.

Following extraction, RNA integrity must be determined, recorded and provided. High quality RNA is essential for RNA-seq library construction. Degraded RNA can lead to increased noise in the resulting mRNA-seq library as measured by transcript coverage, mitochondrial content, ribosomal content and gene diversity. It is recommended that a minimum threshold RIN 7, as measured by an Agilent bioanalyzer, or equivalent, be applied to all REMC RNA samples.

3.3: Protocol: RNA-seq

3.3.1: Amplification

Due to protocol specific biases introduced during RNA or cDNA amplification it is recommended that neither RNA nor full-length cDNA amplification be performed for IHEC RNA-seq libraries.

3.3.2: RNA format and processing

It is recommended that either polyA+ RNA or ribodepleted total RNA be used for IHEC RNA-seq libraries. Both random primed and RNA fragmentation based methodologies are suitable for IHEC RNA-seq library preparation. If a random primed method is used, it is recommended that the subsequent cDNA be rendered single stranded prior to PCR amplification to maintain strand specificity.

3.3.3: Sequencing

Paired-end sequencing should be the method of choice for RNA-seq library sequencing.

3.3.4: Read depth

The number, type and length of sequence reads required to sample a RNA-seq library is dependent on the library preparation methodology and analysis type to be performed. For samples that will form IHEC complete epigenomes, where the goal is to comprehensively represent polyadenylated transcription within a cell or tissue type, a minimum depth of ~200 million paired-end reads per replicate (representing 100 million cDNA fragments) of 75 nucleotides in length is required. In cases where more than two biological replicates of a tissue

or cell type will be performed ~50-100 million paired- end reads per additional replicate (representing 25-50 million cDNA fragments) of 75 nucleotides is sufficient.

3.4: QC metrics: RNA-seq

The following QC metrics should be determined and monitored to ensure that REMC RNA-seq libraries are of high quality:

3.4.1: Exon:Intron ratio

- assessed to detect potential genomic contamination (the mRNA-seq protocol should include a standard DNAse treatment step).

3.4.2: Overall transcript coverage

- assessed to detect mRNA degradation and degree to which full-length mRNA was randomly sampled. Uniform coverage should be obtained for 1Kb transcripts.

3.4.3: Total number of duplicate reads (identical forward and reverse read starts)

- a measure of library diversity. Tissue and cell type dependent but outliers should be investigated as potential library failures.

3.4.4: Fraction of reads mapping to mitochondrial transcripts

- assessed to detect RNA degradation. Tissue and cell type dependent but outliers should be investigated as potential library failures.

3.4.5: Fraction of reads mapping to ribosomal transcripts

- assessed to detect degree of polyA enrichment. Tissue and cell type dependent but outliers should be investigated as potential library failures.

3.4.6: Fraction of reads mapping to intergenic regions

- assessed to detect potential genomic contamination (the RNA-seq protocol should include a standard DNAse treatment step).

3.4.7: Percentage of reads inappropriately aligned to anti-sense strands

This value should be <=1%. This can be determined by enumerating the fraction of exon-exon junction reads aligned to the anti-sense strand. Alternatively known poly-adenylated foreign RNA standards can be spiked into the RNA prior to library construction and used to determine the level of artifact anti-sense transcription.

3.5: QC metrics: miRNA-seq

In parallel to RNA-seq, the small RNA content of a transcriptome can be captured and sequenced (miRNA-seq). miRNA-seq library construction involves; 1) extraction of total RNA; 2) 3' pre-adenylated RNA linker ligation 3) 5' ATP dependent RNA linker ligation; 4) RT using a primer which hybridizes to the 3' RNA linker followed by PCR with primers suitable for solid phase (Illumina) or emPCR (SOLiD) clonal amplification. Following PCR the desired insert size range, typically 18-30 bp, is purified away from ligation and PCR by-products by gel electrophoresis. Taking advantage of the fact that the complexity of miRNA-seq libraries is significantly less than that of RNA-seq libraries, miRNA-seq libraries are typically indexed and pooled prior to sequencing.

The following QC metrics should be determined and monitored to ensure that REMC miRNA-seq libraries are of high quality.

3.5.1: Saturation plots of miRNA species diversity to assess sequence depth

-miRNA diversity appears to be tissue dependent.

3.5.2: Fraction of total reads that are adapter dimer (the major miRNA-seq library construction biproduct)

-Dimer fraction increases as RNA quality and miRNA quantity decrease in a given sample.

3.5.3: Fraction of aligned reads that align to other small RNA species (tRNAs, snoRNAs etc..).

3.5.4: Fraction of aligned reads that align to mature miRNA, miRNA and pre-miRNA sequences.*

4: ChIP-seq

4.1: Introduction

The minimum set of core histone modification includes H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3. Other modifications of core histones and RNA polymerase are required for Class 1 and 2 epigenomes.

4.2: Protocol: ChIP

4.2.1: Chromatin preparation

Isolated cells or tissues may either be fixed when collected, or flash frozen and then subject to fixation upon thawing. Chromatin should be shared by sonication and/or micrococcal nuclease to a sixe range of ~150-500bp. In every sample preparation the chromatin fragment size should be analyzed by Agilent Bioanalyzer or agarose gel electrophoresis prior to commencing with the IP. The condition of fixation and chromatin fragmentation must be optimized for individual tissue/cell samples to achieve a consistent size range.

4.2.2: Antibody

Primary antibodies may be either monoclonal or polyclonal, and should be validated by (i) Western blot to verify specific recognition of histone, and (ii) ELISA or dot blots using modified histone tail peptides to verify specificity for the targeted modification. When antibodies are commercially sourced, the vendor may supply the validation data, provided it is derived from the actual lot in use. A control antibody should be used to validate the ChIP efficiency.

4.2.3: ChIPed material

Whenever applicable, the amount and size of recovered DNA should be assessed by Agilent Bioanalyzer or agarose gel electrophoresis prior to library construction

4.3: Protocol: library construction

Library production should proceed based on manufacturer's protocols, using minimal PCR amplification (recommended range of ~10 cycles with Illumina procedures). Adapters for Illumina sequencing may be either single plex or multiplex ("barcoded"); in the latter case, it is

important to establish that adapters are unique within the pool at a least 2 positions to ensure clean separation of the reads during analysis.

4.4: QC metrics: ChIP-seq

The following QC metrics should be determined and monitored to ensure that IHEC ChIP-seq datasets are of high quality.

4.4.1: Read length and sequencing depth

It is recommended to generate 30-50 million aligned reads with at least a 36 base read length. However, different sequencing depths could be made for different marks based on distribution of coverage for that particular mark. Smaller number of reads (i.e., ~30 million) may be adequate for modifications associated with transcriptional activation, such as H3K4me3 and H3K27ac; however, deeper depth should be considered for broadly distributed marks associated with transcriptional inactivation such as H3K9me3 and H3K27me3.

4.4.2: Fraction aligned reads, duplicate reads.

At present, there are no specific recommendations for these parameters. However, it is evident that these metrics can give important insights into the progress of a Chip-Seq experiment, helping one to detect, for instance, primer dimer artifacts (leading to low fraction aligned reads), contamination of human cell lines by mouse cell lines (same effect), or potential PCR artifacts (leading to a low fraction of unique reads). However, it is essential to take into account the expected complexity of the library (which varies from mark to mark) and the sequencing depth before setting a standard for percent duplicate reads, since over sequencing of a good library can produce perfectly good data but with a reduced fraction of unique reads.

4.4.3: Concordance between replicate datasets.

ChIP-Seq datasets should be replicated at the level of the biological sample, wherever possible. An exception can be made in the case of a tissue type of high clinical significance which is nonetheless difficult to obtain. The replicated datasets should be compared to ascertain consistency and in cases of large discordance and additional replicate generated.

4.4.4: Fraction of reads in enriched intervals, and other criteria.

The Roadmap Epigenomics project has an ongoing effort to define and reduce to practice data based QA criteria. When appropriate, these should be incorporated into IHEC documentation of standards.

4.4.5: Use of controls.

To ascertain the background distribution of fragment abundances in the input material, it is essential to sequence a library derived from the chromatin preparation.

4.5: Biological replicates

At least two biological duplicate should be done (although triplicate is ideal, some tissue samples may be difficult to collect, see above). If the correlation between the replicate is low, another preparation should be made.