Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Genome-wide nucleosome occupancy and DNA methylation profiling of four human cell lines

Aaron L. Statham [a,*], Phillippa C. Taberlay [a,b], Theresa K. Kelly [c], Peter A. Jones [d], Susan J. Clark [a,b]

[a] Epigenetics Research Program, Cancer Division, Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia
[b] St Vincent's Clinical School, Faculty of Medicine, University of New South Wales Australia, Darlinghurst, NSW 2010, Australia
[c] Active Motif, Inc., Carlsbad, CA 92008, USA
[d] Van Andel Research Institute, Grand Rapids, MI 49503, USA

ARTICLE INFO

Article history:
Received 24 November 2014
Accepted 25 November 2014
Available online 8 December 2014

Keywords:
Epigenetics
Methylation
Nucleosome positioning
NOMe-seq

ABSTRACT

DNA methylation and nucleosome positioning are two key mechanisms that contribute to the epigenetic control of gene expression. During carcinogenesis, the expression of many genes is altered alongside extensive changes in the epigenome, with repressed genes often being associated with local DNA hypermethylation and gain of nucleosomes at their promoters. However the spectrum of alterations that occur at distal regulatory regions has not been extensively studied. To address this we used Nucleosome Occupancy and Methylation sequencing (NOMe-seq) to compare the genome-wide DNA methylation and nucleosome occupancy profiles between normal and cancer cell line models of the breast and prostate. Here we describe the bioinformatic pipeline and methods that we developed for the processing and analysis of the NOMe-seq data published by (Taberlay et al., 2014 [1]) and deposited in the Gene Expression Omnibus with accession GSE57498.

Specifications	
Organism/cell line/tissue	Four human cell lines: *HMEC* — normal breast *MCF7* — breast cancer *PrEC* — normal prostate *PC3* — prostate cancer
Sex	*HMEC* and *MCF7* are female, PrEC and *PC3* are male
Sequencer or array type	Illumina HiSeq 2000
Data format	Raw and processed
Experimental factors	Cancer versus normal
Experimental features	Comparison of nucleosome depleted regions between normal vs cancer breast and prostate cell line
Consent	n/a
Sample source location	n/a

Direct link to deposited data

Raw data: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57498.

Processed data: https://zenodo.org/record/12454.

Scripts and manual for analysis: https://github.com/astatham/NOMe-seq-analysis.

Experimental design, materials and methods

The NOMe-seq assay

The Nucleosome Occupancy and Methylation sequencing (NOMe-seq) technique is an extension of whole genome bisulfite sequencing (WGBS) enabling the simultaneous interrogation of nucleosome positioning and DNA methylation on the same strand of DNA [2]. This is achieved through the treatment of isolated cell nuclei with the *M.CviP*I GpC methyltransferase, which adds a GpC methylation "footprint" to accessible regions of the genome. In contrast, endogenous mammalian cytosine methylation occurs almost exclusively in CpG sites. Therefore, accessibility dependent, *M.CviP*I added GpC methylation is separable bioinformatically from endogenous methylation after WGBS by whether the methylcytosines occur within a CpG or GpC sequence context. Cytosines in GpCpG sites are removed from the analysis as they are both a CpG and GpC context, and therefore a mixture of endogenous methylation and accessibility. Additionally *M.CviP*I spuriously methylates CpCpG sites at a low efficiency, so these sites are also discarded. This

[*] Corresponding author at: Epigenetics Research Program, Cancer Division, Garvan Institute of Medical Research, 384 Victoria St Darlinghurst NSW 2010, Australia.

http://dx.doi.org/10.1016/j.gdata.2014.11.012

leaves the human genome (hg19) with approximately 123 million GCH sites available for nucleosome occupancy analysis and 20.4 million WCG sites available for endogenous methylation analysis.

Cell culture

Normal human prostate epithelial cells (PrEC), a prostate cancer cell line (PC3), and a breast cancer cell line (MCF7) were obtained from the American Type Culture Collection (ATCC). Normal human mammary epithelial cells (HMEC) were obtained from Invitrogen. All cell lines were cultured under recommended conditions at 37 °C and 5% CO_2.

Nucleosome occupancy and methylation sequencing (NOMe-seq)

Cells were trypsinized and centrifuged for 3 min at 500 *g*, then washed in ice-cold PBS and resuspended in 1 mL ice-cold Nuclei Buffer (10 mM Tris, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1 mM EDTA, and 0.5% NP-40, plus protease inhibitors) per 5×10^6 cells and incubated on ice for 5 min. Nuclei were recovered by centrifugation at 900 *g* for 3 min and washed in Nuclei Wash Buffer (10 mM Tris, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, and 0.1 mM EDTA containing protease inhibitors). Freshly prepared nuclei (2×10^6 cells) were resuspended in 1× *M.CviP*I reaction buffer (NEB), then treated with 150 units of *M.CviP*I (NEB; 50,000 units/mL) for 15 min in 15 μL 10× reaction buffer, 45 μL 1 M sucrose, and 0.75 μL SAM in a volume of 150 μL. Reactions were quenched by the addition of an equal volume of Stop Solution (20 nM Tris – HCl [pH 7.9], 600 mM NaCl, 1% SDS, 10 mM EDTA, 400 μg/mL Proteinase K) and incubated overnight at 55 °C. DNA was purified by phenol/chloroform extraction and ethanol precipitation.

Libraries for genome-wide NOMe-seq analyses were then prepared using the established protocols of the USC Epigenome Center. Briefly, genomic DNA (2 μg) was sonicated using a Covaris instrument to an average fragment length of 150 bp. Achievement of the desired size range was verified by Bioanalyzer analysis (Agilent Technologies). Fragmented DNA was repaired to generate blunt ends using the END-It kit (Epicentre Biotechnologies) according to the manufacturer's instructions. Following incubation, the treated DNA was purified using AMPure XP beads (Agencourt). Magnetic beads were used for all nucleic acid purifications in subsequent steps. Following end repair, A-tailing was performed using the dA-tailing module according to the manufacturer's instructions (New England Biolabs). Adapters with a 3′ “T” overhang were then ligated to the end-modified DNA. Modified Illumina paired-end (PE) adapters were used. Ligation was carried out using ultrapure, rapid T4 ligase (Enzymatics) according to the manufacturer's instructions. The final product was then purified with magnetic beads to yield an adapter-ligation mix. Prior to bisulfite conversion, bacteriophage lambda DNA that had been through the same library preparation protocol described above to generate adapter-ligation mixes was combined with the genomic sample adapter ligation mix at 0.5% w/w. Adapter-ligation mixes were then bisulfite converted using the Zymo DNA Methylation Gold kit (Zymo Research) according to the manufacturer's recommendations. The final modified product was purified by magnetic beads and eluted in a final volume of 20 μL. Amplification of one-half the adapter-ligated library was performed using KapaHiFi-U Ready Mix under the following conditions: 98 °C for 2 min, followed by four cycles of 98 °C for 30 s, then 65 °C for 15 s and 72 °C for 1 min with a final extension for 10 min in 50 μL total reaction volume. The final library product was examined on the Agilent Bioanalyzer then quantified using the Kapa Biosystems Library Quantification kit according to the manufacturer's instructions. Optimal concentrations to obtain the correct cluster density were determined empirically. Resulting libraries were plated using the Illumina cBot and run on the Illumina HiSeq 2000 platform configured for 100 bp paired-end reads according to the manufacturer's instructions.

Data processing

Raw read pairs were aligned to the human genome (hg19) using bwa-meth [3], a bisulfite aware wrapper for the bwa-mem aligner [4]. Read pairs with identical strand, start and end positions were considered PCR duplicates and removed from downstream analysis using MarkDuplicates from the Picard toolset (http://broadinstitute.github.io/picard/). BisSNP [5] was then used to extract the methylation status of all WCG and GCH sites in the genome for each cell line, and custom scripts (supplied in Supplementary Data 1) then used to tabulate the number of cytosine (i.e. methylated) and thymine (i.e. unmethylated) containing reads across each GCH and WCG in each cell line. These tables were then imported into R for downstream analysis.

Detection of nucleosome depleted regions (NDRs)

A sliding window chi-squared test was used to detect regions of statistically significant increased GCH methylation compared to the whole genome as a background, indicating a potential NDR. The number of methylated (C) and unmethylated (T) counts at GCH sites with at least 5× sequencing coverage was summed in 100 bp windows at 20 bp spacing. Each window was tested against the sum of Cs and Ts of the whole genome, and windows were considered significant when their P-value is $<10^{-15}$. Overlapping significant windows are then joined, and only returned as an NDR if at least 140 bp in size. This procedure is implemented in the ‘findNDRs’ function of the *aaRon* R package (http://github.com/astatham/aaRon) with user-tunable parameters. Fig. 1 shows the number of NDRs found in each cell line by ‘findNDRs’ using the default settings, that many of these NDRs overlap between the four cell lines and that a significant number are also unique to a single cell line.

Visualization of aggregate NOMe-seq profiles

To enable the visualization and comparison of NOMe-seq profiles across sets of genomic regions (for example transcription start sites or transcription factor binding sites) the ‘methylationPlotRegions’ family of functions was implemented in *aaRon*. Fig. 2 shows average profiles of both accessibility and DNA methylation surrounding transcription start sites (TSSs) in MCF7 cells that do and do not overlap a H3K4me3 ChIP-seq peak, an active promoter chromatin mark. As expected, the active, H3K4me3 marked TSSs show a distinct highly accessible region

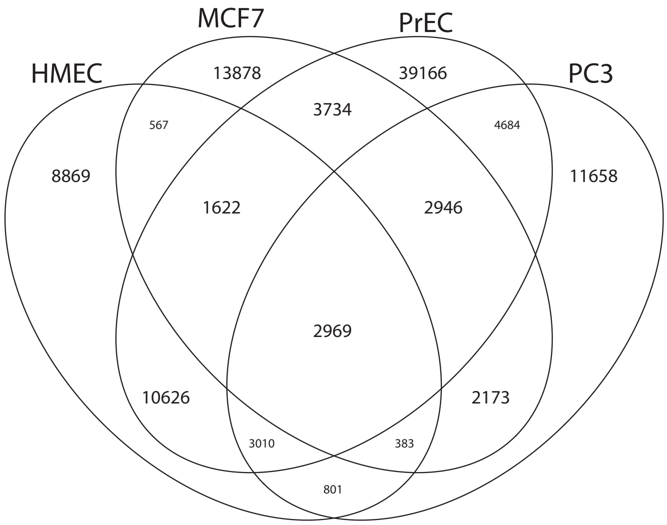

Fig. 1. Venn diagram of overlap between nucleosome depleted regions found in the four cell lines. Statistically significant NDRs calls were obtained from the ‘findNDRs’ function and overlaps between the cell lines were plotted with the ‘grangesVenn’ function in *aaRon*.

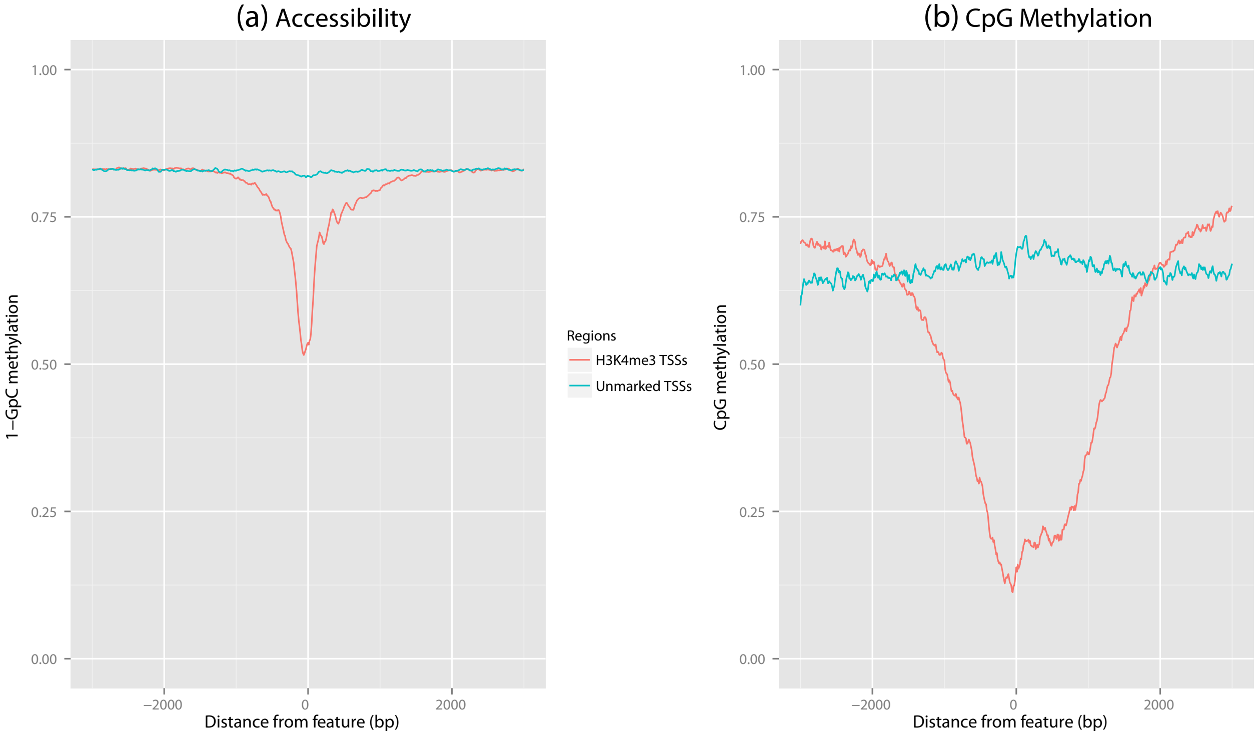

Fig. 2. MCF7 aggregate NOMe-seq profiles of the (a) accessibility and (b) CpG methylation levels surrounding transcription start sites split by the presence and absence of the H3K4me3 chromatin mark. UCSC knownGene TSSs were overlapped with ENCODE MCF7 H3K4me3 ChIP-seq peaks obtained using *rtracklayer* and *AnnotationHub* Bioconductor packages respectively. NOMe-seq profiles of the H3K4me3 and unmarked TSSs were plotted against each other using the 'methylationPlotRegions' function in *aaRon*.

at the TSS with strongly phased nucleosomes downstream whereas the inactive, unmarked TSSs are completely inaccessible (Fig. 2a). In addition, DNA methylation is depleted in a multi-kilobase region surrounding H3K4me3 marked TSSs, in contrast to the consistent high methylation level across unmarked TSSs (Fig. 2b). Examples of other visualizations of NOMe-seq data using *aaRon* are included in Supplementary Data 1.

Discussion

Here we describe the bioinformatic pipeline and tools we have developed for the analysis and visualization of NOMe-seq data for four human cell lines [1]. The step-by-step manual supplied in Supplementary Data 1 enables other researchers to precisely reproduce and build upon the analysis methods we have created, fostering advancement of the epigenomics and bioinformatics fields as a whole.

Software availability

All software, the exact commands and custom scripts used for this analysis have been documented in detail and are contained in Supplementary Information 1 and are also available from https://github.com/astatham/NOMe-seq-analysis. The 'findNDRs', 'grangesVenn', 'methylationPlotRegion' functions and others described in Supplementary Information 1 are part of the *aaRon* R package available from http://github.com/astatham/aaRon.

Acknowledgements

This work was funded by National Health and Medical Research Council Grant 1051757.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.11.012.

References

[1] P.C. Taberlay, A.L. Statham, T.K. Kelly, S.J. Clark, P.A. Jones, Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 24 (9) (2014) 1421–1432.

[2] T.K. Kelly, Y. Lie, F.D. Lay, G. Liang, B.P. Berman, P.A. Jones, Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. 22 (12) (2012) 2497–2506.

[3] B.S. Pedersen, K. Eyring, S. De, I.V. Yang, D.A. Schwartz, Fast and accurate alignment of long bisulfite-seq reads. arXiv:1401.1129 [q.bio.GN] (2014).

[4] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN] (2013).

[5] Y. Liu, K.D. Siegmund, P.W. Laird, B.P. Berman, Combined DNA methylation and SNP calling for bisulfite-seq data. Genome Biol. 13 (7) (2012) R61.