Identification of Chromatin Accessibility from Nucleosome Occupancy and Methylome Sequencing

Yongjun Piao¹, Seongkeon Lee², Keith D. Robertson³, Huidong Shi⁴, Keun Ho Ryu⁵, Jaeng-Hyeon Choi⁶

¹College of Electrical and Computer Engineering, Chungbuk National University, South Korea
²Department of Statistics, Sungshin Women’s University, South Korea
³Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA
⁴Department of Biochemistry and Molecular Biology, Georgia Regents University, Augusta, GA, USA
⁵Department of Biostatistics and Epidemiology, Georgia Regents University, Augusta, GA, USA

December 11, 2015
As you know, human genome contains 3 billion base pairs of DNA packed into 23 chromosomes. Since each base pair is around 0.34 nanometers long, so it is estimated that the human body contains about 50 trillion cells—which works out to 100 trillion meters of DNA per human. So the question is how exactly does the cell actually take such a long molecular and place it into the small nucleus of cells.

Basically, our double helix DNA is wrapped around with specific proteins, known as histones, with the structure of 8 histones are wrapped with 147 base pair of DNA to form the nucleosomes. Then the nucleosomes fold up to form tube-like fiber, known as chromatin.

Various studies have been show that the chromatin accessibility plays a key role in epigenetic regulation of gene activation and silencing. In other words, open chromatin regions (nucleosome-depleted regions; NDRs) allow regulatory elements such as transcription factors and polymerases to bind for gene expression while closed chromatin regions (nucleosome-occupied regions; NORs) prevent the activity of transcriptional machinery.
Recently, Nucleosome occupancy and methylome sequencing, short as NOMe-seq has been developed to allow simultaneously profiling chromatin accessibility and DNA methylation on single molecules.

The NOMe-Seq assay works by treating the chromatin with a GpC methyltransferase enzyme to artificially methylate GpC dinucleotides that are not protected by nucleosomes or other proteins bound to the DNA. Treated DNA is then subjected to bisulfite conversion and are sequenced to establish the DNA methylation profile for a single DNA strand.

However, to our knowledge, there is no standard method for de novo identification of chromatin accessibility from NOMe-seq.
So, in this work, we developed a computational method to identify chromatin accessibility from NOMe-seq experiments.

First the raw reads were mapped to a reference genome and the methylation scores of all GCHs are quantificated. The methylation score beta is calculated by methylated reads divided by total number of methylated and unmethylated reads.

Then, scanning the genome from 5’ to 3’ direction, it detects seeds, we defined seed as a GCH is a seed if its the methylation score is smaller than pre-defined threshold. Then, extends the seed based on different stopping conditions.
Let’s see an example.
Suppose that we have 11 GCHs, the methylation scores of them are indicated in the figure. Delta is the seed detection threshold, which was set to 0.1, and mu is the average methylation score threshold 0.2, big delta is the jump threshold. So, in this example, we have only two GCHs: GCH1 and GCH2 to be the seeds, since their methylation score is smaller than 0.1.
Then, we extend the seed region by examining the nearest peak and valleys. If the methylation score of a peak is larger the big delta, it will stop extension, but in this case the score is 0.45, thus we check the average methylation score of all GCHs from seed to valley.
Its 0.193 which is smaller than the average methylation score, So we continue extension by searching next peak and valley. However, the methylation score of the next peak is 0.61, so the extension should be stopped.

But if we stopped here, the detected region will be GCH1 to 6, it’s not a optimal solution. So we used non-parametric mixture models to decide exact end point.
To evaluate the efficiency, we tested our method both on simulated and read datasets. Since the true nucleosome positions are unknown in human genomes, we proposed a simulation model.

These are distributions of four simulated data generated by different simulation parameters.

(MuO and sigmaO indicates the mean and standard deviation of GCHs within NOR, muD and sigmaD refers to the mean and standard deviation of GCHs within NDR.)

X-axis shows the methylation score, and the red line indicates the methylation score distribution of GCHs within NOR, the blue line indicates the methylation score distribution of GCHs within NDR.

So, as shown in the figure a, the methylation score of GCHs in NORs and NDRs are well-separated, and for figure b, c, d, we changed the simulation parameters to make NORs and NDRs be much more overlap with each other.
Then, we performed ROC analysis and compared the results with CpG_MPs and Max-gap-min run.

Typically, CpG_MPs was designed for identifying DMRs, but we can easily apply it on NOMe-seq. and Max-gap-min-run is a commonly used segmentation algorithm. So we compared the performance with these two methods.

The black indicates our results, the pink one refers to the result of maxgapminrun, the other one indicates CpG_MPs. The roc curves show that our method outperforms the other two%. 
Until now, we evaluated our method by individual GCHs, here we investigated the results by regions.

We examined that how much NORs predicted by our method overlap the actual NORs. The x-axis indicates the overlapping fraction, so if the fraction is 1, it means predicted NORs are completely overlapped with actual NORs. Y-axis refers to the frequency. Blue one is ours, green is CpG_MPs, and the other one is max-gap-min-run’s result. As you can see in the figures, our method has larger number of NORs completely overlapped with actual NORs compared with CpG_MPs and Max-gap-min-run. Moreover, CpG_MPs and max-gap-min-run have some regions fraction smaller than 0.2, which mean they cannot detect some NORs, while ours are good.

These figures are just like opposite to the previous figures. At this time, we looked how much predicted NORs overlap the actual NDRs. So in this case, when the fraction is 1, it means predicted NORs are completely overlapped with actual NDRs, thus it made a wrong prediction to NDRs.

As shown in the figure, our method has much more accurate prediction than the other two methods. And CpG_MPs and max-gap-min-run has large number of wrong predictions. We found that these two methods sten
d to merge too much regions.
Then we tested our method on hct116 wild type data.

These are the summary of the data. It has about 2 million GCHs in chromosome1 in the control data, and has about 4 million GCHs in M.CviPI treated data. Among them, about 1 million GCHs exists in both control and enzyme treated data. So we only considered these 1 million GCHs.

To compare with the region found by our method, we constructed actual NORs and NDRs as follows. So if the methylation score differences between M.CviPI treated GCHs and control GCHs is smaller than 0.2, we considered them as unmethylated, so belong to NORs. Otherwise, treated as methylated, so belong to NDRs.

So these figures show the methylation score distributions of the gold standard. The left figure is the actual distribution, but it has too many GCHs with 0 scores. To better look into it, I removed the GCHS with 0 scores and re-drew the distribution. We can see that the methylation score of GCHs within the NOR and NDR are reasonably distributed, although there are some exceptions in here, here, and here.
And then we also compared the results with CpG MPs and max-gap-min-run. CpG MPs has competitive sensitivity compared with our method, but the specificity is extremely low. As I said, this is the direct result that CpG MPs tends to merged too many regions. The max-gap-min-run also had poor results.
Similar result also can be seen in evaluation of regions.
So, our goal is to identify cancer heterogeneity in DNA methylation studies

---

**Summary**

- NOMe-seq is an innovative technology to measure DNA methylation and nucleosome occupancy simultaneously
- Novel algorithm was presented for identifying chromatin accessibility
  - Identify seeds which are very likely GCHs in NORs
  - Extend seed as long as the average of GCH methylation score is smaller than a threshold
  - Decide the endpoint based mixture model
- Proposed method has extremely high accuracy
Acknowledgment

Lee Lab at SWU

Shi Lab at GRU
Huidong Shi
Eun-Jun Lee
Leah Pei
Jinfeng Luo
Jinmei Liu
James Wilson
Ethan Sperti

Collaborators
John Cowell (GRU)
Ahmed Chadli (GRU)
Hanlio Ding (GRU)
Yanbin Dong (GRU)
Zheng Dong (GRU)
Vadivel Ganapathy (GRU)
Zhonglin Hao (GRU)
Larry Laymen (GRU)
Kaitlin Liu (GRU)
Naufal Mвеghi (GRU)
Betty Paxo (GRU)
Shuyong Su (GRU)
Mohammed Thangaraju (GRU)
Dorothy Tuen (GRU)
Xiaolong Wang (GRU)
Huidong Zhu (GRU)
Hanu Tang (IJ)
Clev Fous (II)
Ellen Wietzner (II)
John Colbourne (II)
Jack Warren (IJ)

Grants Support
NIH