Chromatyping: Reconstructing Nucleosome Profiles from NOMe Sequencing Data

Shounak Chakraborty[1,2,3,4], Stefan Canzar[4], Tobias Marschall[2,3(✉)], and Marcel H. Schulz[1,2,3(✉)]

[1] Cluster of Excellence for Multimodal Computing and Interaction, Saarland University, Saarland Informatics Campus E1.7, 66123 Saarbrücken, Germany
mschulz@mmci.uni-saarland.de
[2] Max Planck Institute for Informatics, Saarland Informatics Campus E1.4, 66123 Saarbrücken, Germany
t.marschall@mpi-inf.mpg.de
[3] Center for Bioinformatics, Saarland University, Saarland Informatics Campus E2.1, 66123 Saarbrücken, Germany
[4] Gene Center, Ludwig-Maximilians-Universität München, 81377 Munich, Germany

Abstract. Measuring nucleosome positioning in cells is crucial for the analysis of epigenetic gene regulation. Reconstruction of nucleosome profiles of individual cells or subpopulations of cells remains challenging because most genome-wide assays measure nucleosome positioning and DNA accessibility for thousands of cells using bulk sequencing. Here we use characteristics of the NOMe-sequencing assay to derive a new approach, called ChromaClique, for deconvolution of different nucleosome profiles (chromatypes) from cell subpopulations of one NOMe-seq measurement. ChromaClique uses a maximal clique enumeration algorithm on a newly defined NOMe read graph that is able to group reads according to their nucleosome profiles. We show that the edge probabilities of that graph can be efficiently computed using Hidden Markov Models. We demonstrate using simulated data that ChromaClique is more accurate than a related method and scales favorably, allowing genome-wide analyses of chromatypes in cell subpopulations. Software is available at https://github.com/shounak1990/ChromaClique under MIT license.

Keywords: NOMe-seq · Max clique enumeration · Epigenetics HMMs

1 Introduction

The eukaryotic genome is organized in nucleosomes which consist of approximately 147 base pairs of DNA wrapped around a histone octamer. Nucleosomes serve as the basic unit of chromatin packaging and are connected via free DNA

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 21–36, 2018.
https://doi.org/10.1007/978-3-319-89929-9_2

linkers of variable length. Nucleosome positioning plays a pivotal role for transcriptional regulation by controlling DNA accessibility for binding proteins (*e.g.* transcription factors). Thus, learning more about nucleosome positioning and how it differs between different cell types, as well as subpopulations of cells, is an important task to understand gene expression regulation.

Different protocols for the genome-wide characterization of nucleosome positioning have been developed. The most common are DNaseI-seq [1], ATAC-seq [2] and NOMe-seq [3]. NOMe-seq (nucleosome occupancy and methylation) utilizes the enzyme M.CviPI which specifically methylates cytosine dyads in a GpC sequence context. Because NOMe-seq uses bisulfite sequencing, it also delivers the endogenous CpG methylation levels, enabling the simultaneous analysis of chromatin accessibility and DNA methylation. Due to this unique feature, a number of recent studies have applied NOMe-seq to study epigenetic regulation [3–7]. It is also the first assay that can measure nucleosome positioning and DNA methylation simultaneously in single cells [8].

However, single cell datasets using NOMe-seq or other related assays are rare, whereas bulk sequencing experiments do not reveal nucleosome and chromatin profiles of subpopulations of cells. Although NOMe-seq is normally obtained from bulk sequencing of cells, the nucleosome readout of one paired-end read comes from a single cell. As several GpC dinucleotides may appear on a paired-end read obtained from NOMe-seq, this information can be used to group reads that originate from the same nucleosome profile. We call these distinct nucleosome profiles *chromatypes*, to emphasize that their chromatin arrangement differs between cells. Here, we are concerned with the development of novel computational methods that can reconstruct chromatypes from NOMe-seq data.

The only comparable method is epiG, which clusters reads according to epigenetic haplotypes using a Bayesian approach that considers DNA methylation and GpC methylation in NOMe-seq data [9]. However, the Bayesian approach in epiG is slow and can thus only be used to study local genomic regions and does not allow genome-wide application.

We exploit recent advances for methods that reconstruct viral haplotypes from DNA-seq data. The high mutation rates of viruses such as HIV give rise to considerable intra-patient variability of virus genomes [10]. Reconstructing the full set of virus haplotypes ciruclating in a patient's blood and quantifying their relative abundances are important tasks with the prospect of informing therapy stratification [11]. This computational task is challenging, however, because usually no *a priori* knowledge on the number of haplotypes and the distribution of their abundances is available. Therefore, distinguishing sequencing errors from low-abundance haplotypes requires non-trivial techniques. In the meantime, a wealth of methods has been developed [12], including Haploclique that enumerates maximal cliques on a DNA-seq read graph [13].

We introduce a novel method, called ChromaClique, which combines the maximal-clique enumeration procedure of HaploClique with a novel probabilistic edge criterion tailored to NOMe-seq data. The edge criterion incorporates base quality scores in a probabilistic manner. ChromaClique uses Hidden Markov Models for the efficient computation of the edge probabilities in the novel read

graph. We show that ChromaClique is the first algorithm that can be used genome-wide and that it has better accuracy on simulated data compared to the only comparable method epiG.

2 Methods

2.1 ChromaClique Overview

ChromaClique starts from bulk NOMe-seq reads aligned to a reference genome in BAM format. Each cell, or group of cells, is expected to have different nucleosome positioning patterns (chromatypes) which are encoded in the reads. This is depicted in Fig. 1 with the different colors, where each color represents a chromatype. The aligned reads are converted into a *read graph*, $G := (V, E)$, with nodes V and edges E. Each node represents a read. Two reads share an edge only if they are likely to originate from the same chromatype. Both single and paired-end reads are considered for the edge criterion. Two paired-end reads share an edge only when both reads from both pairs agree to the edge criterion. The maximal cliques in the graph are enumerated using the algorithm previously employed in HaploClique [13]. The reads in a maximal clique are merged. The condensed graph is checked again for cliques which have an edge between each other and the maximal clique finding algorithm is run iteratively. This continues until no more edges are found in the graph. The nodes in the final graph represent the individual reconstructed chromatypes and are also called super reads.

2.2 Encoding the Reads

In NOMe-seq data only GCH trinucleotides, *i.e.* GCT, GCA or GCC, in the genome provide information about open and closed nucleosome positions,

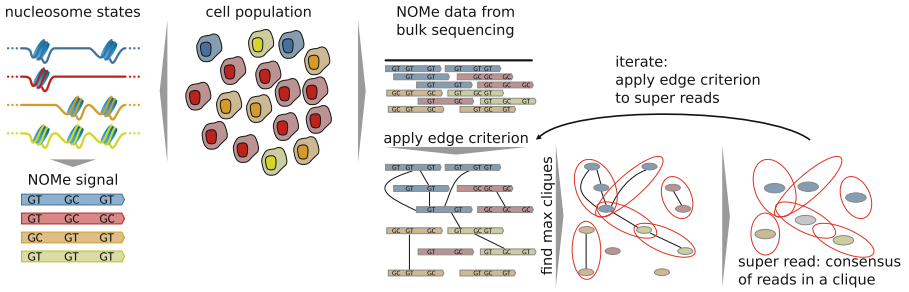

Fig. 1. Illustration of a cell population with different nucleosome states, indicated by different colors. The NOMe signature of different chromatin states is shown on the bottom left. The ChromaClique workflow is shown on the right: ChromaClique applies its edge criterion to NOMe bulk sequencing data (black lines connecting reads), enumerates all maximal cliques (indicated in red), merges reads in a clique, and iterates the process until convergence. (Color figure online)

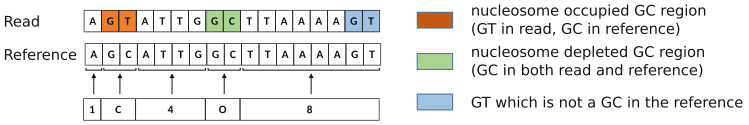

Fig. 2. GCH and GTH positions in a read are compared with the reference. If there is a position with GCH in the read and a GCH in the reference, it is marked as "O" and otherwise if it is a GTH in the read it is marked as "C". All other positions other than the GCH or GTH positions are summarized by numbers reflecting their length. If there is a GTH in the read and a GTH in the reference it is not treated as a GCH position.

because GCG positions are ambiguous due to the possibility that CpG DNA methylation took place. Each individual read is represented as follows: each GCH position is encoded as open (O) and each GTH position is encoded as closed (C). This is because a GCH is not converted by bisulfite treatment, when it is accessible and was methylated by the enzyme. The NOMe enzyme M.cviPI works on accessible GCs on both DNA strands in 5' to 3' direction. This reverse complementarity is taken into account during read encoding and edge construction by the algorithm. Figure 2 shows the process of encoding a read based on its GCH and GTH occurrences. The last GTH position is not a closed position since the reference does not have a GCH position there.

Sequencing errors in the reads that would prevent the detection of GCH can be corrected by comparing the positions in the reference sequence to which the read is aligned. For example if there is a GCH position in the reference and due to an error the read contains HCH instead this will be corrected for later use in the algorithm.

2.3 Definition of Edge Probabilities

In order to build the graph for finding maximum cliques, each pair of reads with sufficient overlap is scored against each other to see if they are likely to originate from cells with the same chromatype. The reads are scored on the basis of their base quality scores (Phred scores) as reported by the sequencer, and also based on the similarity of nucleotides observed at their shared GCH positions.

ChromaClique does not make an assumption on the number and relative abundances of open-chromatin patterns. In order to evaluate the likelihood of an edge between two considered reads, we compute the *edge probability* as the probability that the overlapping portion of both reads has been generated by any one of the possible chromatypes. Before we can properly define the edge probability we make a number of definitions.

ChromaClique first encodes reads at only GCH positions (in the following called C/O positions) and records the distance between consecutive occurrences (see Fig. 2). In the read each GC is denoted as open (O) and each GT as closed (C). For simplicity, we denote $C_i(R)$ as the open-chromatin status (O or C) at the i_{th} C/O position in read R, e.g., $C_2(R) = O$ in Fig. 2. The phred base

quality of the Cytosine or Thymine at the i_{th} C/O position in read R is denoted $phred(i, R)$. Let $Q_i(R)$ be the scaled base quality score at position i, that is $Q_i(R) = 10^{-\frac{phred(i,R)}{10}}$. The distance between the i_{th} and j_{th} C/O position in the read is given by $d_{i,j}(R)$, e.g., $d_{1,2}(R) = 4$ in Fig. 2.

Computing the edge probability involves two steps. The first estimates the probability for a given chromatype y given the base qualities obtained from the sequencer, denoted $P(R|y)$. Let T be the total number of C/O positions in an encoded read, then:

$$P(R|y) = \prod_{i=1}^{T} f_{qual}(R, y, i), \tag{1}$$

where f_{qual} is defined as:

$$f_{qual}(R, y, i) = \begin{cases} 1 - Q_i(R), & \text{if } C_i(R) = C_i(y) \\ Q_i(R), & \text{if } C_i(R) \neq C_i(y) \end{cases}. \tag{2}$$

The second step consists in computing the probability of an individual chromatype y, denoted as $P(y)$. A nucleosome occupies around 147 bps and therefore not all possible chromatypes are equally likely. For example 1 C 2C 2C 2C is more likely than 1 C 2O 2C 2O. We capture this by defining transition events at adjacent C/O positions.

For a read R a *transition* for position i is defined as $C_i(R) \neq C_{i+1}(R)$, namely the open-chromatin state at position i has changed compared to its adjacent position i+1. Here we do not distinguish the direction of the transition, i.e. a transition from an O to a C is equivalent to a transition of a C to an O. Similarly, position i is called a *non-transition* if $C_i(R) = C_{i+1}(R)$. As mentioned above, the distance d between two positions i and j should influence the likelihood of a transition event. Therefore we obtain the empirical transition probability $tr(d)$, as the relative frequency of transition events for a certain distance d:

$$tr(d) = \frac{Transition(d)}{Transition(d) + NonTransition(d)}, \tag{3}$$

where $Transition(d)$ and $NonTransition(d)$ are the number of transition and non-transition events at distance d observed in all reads, respectively. Then the non-transition probability is simply given by:

$$1 - tr(d). \tag{4}$$

Transition or non-transition probabilities are used in the computation of observing a certain C/O pattern in a read. In addition, these probabilities may help to recognize errors in the reads, for instance errors due to the incorrect methylation of the M.CviPI enzyme, or due to incorrect bisulfite conversion. For example if the transition probability for a specific distance, say 10, is 0.05, it means that the number of non-transitions seen for this distance is much higher than the number of transitions. However, if a transition was observed at this distance, the probability that it is an error due to either a failed NOMe or

bisulfite conversion, would be high. This information is later used as a prior when two reads are compared to see if they originate from cells with similar chromatypes.

Finally, we can use the transition probabilities (Eq. 3) to quantify the probability of observing a particular chromatype y. We define:

$$P(y) = \prod_{i=1}^{T-1} f_{transition}(y, i), \tag{5}$$

$$f_{transition}(y,i) = \begin{cases} 1 - tr(d_{i-1,i}(y)), & \text{if } C_i(y) = C_{i-1}(y) \text{ and } i > 1 \\ tr(d_{i-1,i}(y)) & \text{if } C_i(y) \neq C_{i-1}(y) \text{ and } i > 1\,. \\ 0.5 & \text{i} = 1 \end{cases} \tag{6}$$

Intuitively, $P(y)$ will be low if the chromatin state configuration in y is unlikely given the transition probabilities. If two reads R_1 and R_2 are independent of each other, the probability that they originate from a particular chromatype y can now be calculated as follows:

$$P(R_1, R_2|y) = P(R_1|y)\ P(R_2|y). \tag{7}$$

From the law of total probability, the probability that two reads originate from the same chromatype can be computed as:

$$P(R_1, R_2) = \sum_{y \in Y} P(R_1, R_2|y)\ P(y), \tag{8}$$

where Y is the set of all possible 2^T chromatypes. Equation (8) is the central *edge probability* of ChromaClique that is used for building its read graph. Two reads are said to be from the same chromatype if the probability $P(R_1, R_2)$ is above a threshold δ. We call δ the *edge threshold* and only edges with $P(R_1, R_2) > \delta$ are considered in the read graph. δ needs to be set manually by the user, but we will determine a practical value for δ using simulations.

Minimum Overlap. The edge probability depends on another parameter which also needs to be set manually. It is the number of C/O positions, D, in the overlapping portion of the two reads in question. If D is too small then this may lead to false edges between reads originating from different chromatypes. However if the number is too large then it leads to many read overlaps not being considered. By default we set the minimum number of overlapping C/O positions to 2.

Thus this parameter determines the purity of the cliques and also the length of the final super reads. It was set manually after analysing the behaviour of simulated data.

2.4 Efficient Calculation of Edge Probabilities in ChromaClique

The probability that two reads originate from the same chromatype is given by Eq. (8). In order to obtain the above probability R_1 and R_2 have to be

checked against all the possible chromatypes, i.e the entire set Y. The size of Y is 2^T, where T is the number of C/O positions in the overlapping portion of the reads. Thus, it becomes computationally expensive to enumerate all the different chromatypes and then calculate the probability.

However, if the overlapping portion of the reads is modeled as a Hidden Markov Model (HMM), the forward algorithm can be used to efficiently calculate the entire probability without having to enumerate all the possible chromatypes.

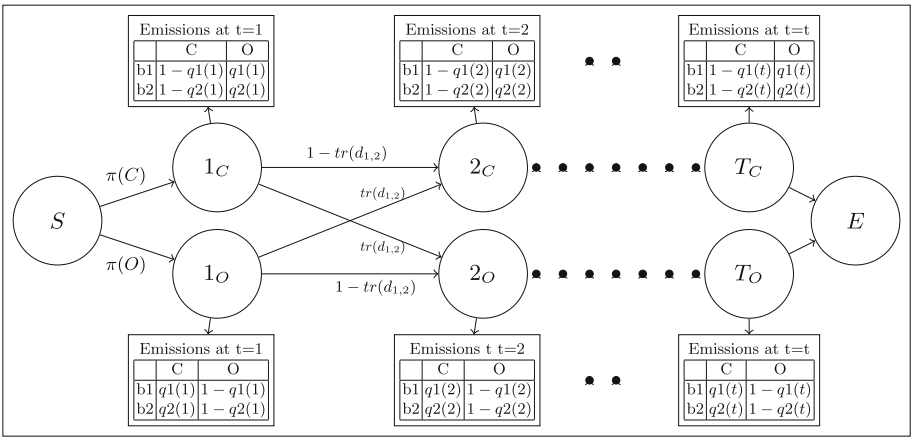

Fig. 3. Hidden Markov Model for calculating the probability that two reads originate from the same chromatype. The circles (1_C, 1_O, 2_C...) represent the hidden states which are the actual open or closed state of the chromatin in the DNA sequence (based on GCH or GTH positions). Each of these states emits two values (one for each of the sequences being compared). The emission probabilities for these are given by the tables near these states. The transition probabilities from one hidden state to another is given by the arrows between the hidden states. The start state S and end state E are customary states denoting the start and end of the process.

Hidden Markov Model (HMM) for Chromatyping. Figure 3 illustrates the HMM for calculating the probability that two reads originate from the same chromatype. It consists of a set of hidden states, which represent the actual nucleotide state (open or closed). Each hidden state emits a pair of nucleotides, one nucleotide for each read at a C/O position. The emission parameters consider the phred base qualities.

More formally, let T be the total number of C/O positions in the overlapping region of the two reads (R_1 and R_2) being compared. Let $t \in \{1, \ldots, T\}$ be the index for the C/O positions, where $R_1(t) \in \{C, O\}$ and $R_2(t) \in \{C, O\}$ denote the chromatin status given by R_1 and R_2 at position t, respectively. Let $\{S, 1_C, 1_O, 2_C, 2_O, ..., T_C, T_O, E\}$ represent the set of hidden states, where S and E denote the silent *start* and *end* state, respectively. In the following we will refer to a state from the set as t_b with $t \in \{1, \ldots, T\}$ and $b \in \{C, O\}$.

State t_b has emission probability $e_{t_b}(b_1, b_2)$ for a pair (b_1, b_2), with $b_i \in \{C, O\}$, defined as:

$$e_{t_b}(b_1, b_2) = \begin{cases} (1 - q_1(t)) \cdot (1 - q_2(t)) & b = b_1 \text{ and } b = b_2, \\ q_1(t) \cdot (1 - q_2(t)) & b \neq b_1 \text{ and } b = b_2, \\ (1 - q_1(t)) \cdot q_2(t) & b = b_1 \text{ and } b \neq b_2, \\ q_1(t) \cdot q_2(t) & b \neq b_1 \text{ and } b \neq b_2, \end{cases} \tag{9}$$

where $q_1(t)$ is defined as:

$$q_1(t) = 10^{-\frac{phred(t,R_1)}{10}} \tag{10}$$

and $q_2(t)$ is defined analogously for R_2.

The initial probabilities from the start state S to 1_C and 1_O are set to $\pi(C) = 0.5$ and $\pi(O) = 0.5$, respectively. The transition probabilities between consecutive states $(t-1)_b$ and t_c, with $b, c \in \{C, O\}$, are defined using the transition probability $tr(d)$ for distance d between C/O positions $t-1$ and t:

$$a_{(t-1)_b, t_c} = \begin{cases} 1 - tr(d_{t-1,t}) & b = c, \\ tr(d_{t-1,t}) & b \neq c. \end{cases} \tag{11}$$

We can now compute the sought probability $P(R_1, R_2)$, Eq. (8), using the standard forward algorithm for HMMs [14]. The complexity of calculating the probability of two reads originating from one chromatype using the forward algorithm is $\mathcal{O}(T)$, where T is the number of C/O positions in the overlapping portion of the reads.

3 Data Simulation and Evaluation

To assess performance with respect to ground truth chromatypes, which are usually not available for real data, we simulated NOMe sequencing experiments *in silico*. Simulated data also serve to tune parameters as needed, in particular δ, the threshold for the probability that two reads originate from cells with the same chromatypes, and **D**, the minimum number of C/O positions we require in the overlapping region of two reads.

3.1 Simulating Chromatypes

The reference sequence of human chromosome 1 was randomly annotated with regions of open chromatin and closed chromatin. Regions of 177 bps were annotated with a nucleosome (closed chromatin for 147 bps) followed by a linker DNA (open chromatin for 30 bps) with a 60% probability. The whole region (177 bps) was annotated as being open chromatin with a 40% probability. This process of annotation was done along the complete chromosome 1. The process was repeated four times in order to simulate four different chromatypes.

Virtual NOMe and bisulfite treatment was simulated as follows: GCHs in nucleosome occupied regions were converted to GTHs. In regions not occupied

by nucleosomes and in linker DNA regions, GCHs were retained. We randomly methylated HCGs, i.e., sites of DNA methylation. In this way each chromatype had distinct open chromatin (GCHs) and DNA methylation (HCGs) profiles, where DNA methylation values are currently only used by epiG.

3.2 Simulating NGS Reads

Illumina sequencing reads were simulated (along with sequencing errors), individually for each of the simulated chromatypes using the ART software [15] and subsequently merged using samtools. The merged reads were aligned to the reference using BISMARK [16]. Four different sets of merged reads, 100 bp reads at 40× and 80× coverage, as well as 200 bp reads at 40× and 80× coverage, were created. We chose 100 bp reads since this is a common read length, while 200 bp reads were included to evaluate the impact of read length on performance. ChromaClique and epiG were run individually on each of these datasets.

3.3 Evaluation Metric for Chromatype Reconstructions

The chromatype reconstructions produced by the algorithms were evaluated based on the number of switches needed to reconstruct that particular super read from the four ground truth simulated chromatypes. Each super read (chromatype-reconstruction) was represented by a binary vector, $Sr[x]$, containing 1 s and 0s, for open and closed positions, respectively.

For example, let S = 1O 42C 23C 9C be a reconstructed super read. This super read can be represented as a binary vector Sr containing 1 s and 0 s for open and closed positions respectively, $Sr = [1, 0, 0, 0]$.

Because the super reads are aligned to the reference, similar vectors can be constructed for each of the ground truth chromatypes that were used for simulating the data. This produces a chromatype matrix $Chr[c, x]$, where each row c represents one of the ground truth chromatypes and each column x represents the nucleosome state (1 or 0) at that position.

For example assume the following chromatype matrix Chr:

$$Chr[c,x] = \begin{matrix} chromatype1 \\ chromatype2 \\ chromatype3 \\ chromatype4 \end{matrix} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}. \tag{12}$$

The number of switches (jumps from one original chromatype to another) required to recreate a particular super read (read group in case of epiG), is referred to as the *switch error*. SE_i is the switch error for super read i. The switch error can be efficiently calculated from the $Sr[x]$ vector and the $Chr[c, x]$ matrix using dynamic programming.

With $SE[c, x]$ we denote the *switch error matrix*, where a row c represents one of the initial chromatypes and an entry in column x denotes the minimum number of switches and mismatches needed to reconstruct a prefix of length x in

the super read starting from the initial chromatype c. $SE[c, x]$ can be formally defined as follows:

$$SE[c,x] = \min \begin{cases} 0, & \text{if } x = 0 \\ SE[c', x-1] + mismatch(c,x), & \text{with } c = c' \text{ and } x \neq 0 \\ SE[c', x-1] + 1 + mismatch(c,x) & \text{with } c \neq c' \text{ and } x \neq 0, \end{cases} \tag{13}$$

where the case $x = 0$ serves as initialization for the DP computation. The function $mismatch(c, x)$ returns 0 if the open chromatin value at position x is same for the Sr vector and the row c in the corresponding chromatype matrix Chr. $mismatch(c, x)$ is defined as:

$$mismatch(c,x) = \begin{cases} 0, & \text{if } Sr[x] = Chr[c,x] \\ 1, & \text{if } Sr[x] \neq Chr[c,x] \end{cases}. \tag{14}$$

The minimum element in the last column of the $SE[c, x]$ matrix represents the minimum number of switches and mismatches required to reconstruct the super read S from any of the given chromatypes. This is defined as the *switch error* of the given super read.

Example Calculation. Recursive application of Eq. (13) on Sr and Chr yields the following SE matrix:

$$SE[c,x] = \begin{array}{l} \\ chromatype1 \\ chromatype2 \\ chromatype3 \\ chromatype4 \end{array} \begin{array}{c} \begin{array}{ccccc} x & 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 1 & 2 \\ 0 & 1 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 & 2 \end{pmatrix} \end{array}. \tag{15}$$

The minimum number of switches and mismatches needed to reconstruct the super read S from the initial chromatypes in this example is 1.

Switch errors were calculated for each super read and the total prediction error is then calculated as an average of the number of switch errors per C/O position as follows:

$$PredictionError = \frac{\sum_i^N SE_i}{\sum_i^N T_i}, \tag{16}$$

where N is the total number of super reads and T_i is the total number of C/O positions in the i_{th} super read.

3.4 Evaluation of the Chromatype Predictions

The super reads produced by ChromaClique were compared against the predictions from epiG and also a *BaseLine chromatype* containing only closed positions. This section talks about the evaluation of the predictions of the different algorithms.

Evaluation of the Output from ChromaClique. ChromaClique outputs a BAM file containing both paired-end and single-end super reads, which are aligned to the reference. Each super read represents local reconstructions of a chromatype. For a single-end super read, the *Chr* matrix and *Sr* vectors can be directly constructed from the nucleotide positions (open or closed), in the super read and the initial chromatypes used for simulation. Thus, the switch error can be calculated directly.

However, for paired-end reads, there is missing information in between the two read ends and the *Chr* matrix needs to be constructed for an individual pair. Essentially, only positions that are overlapped by one of the reads in the super read pair are part of the corresponding *Chr* matrix for that super read, ignoring C/O positions in the reference that are not overlapped by the super read.

Evaluation of the Output from epiG. The output from epiG is not exactly the same as that from ChromaClique. While ChromaClique reports reconstructed local chromatypes obtained by merging reads from the initial aligned reads, epiG assigns reads to "epigenetic haplotypes" [9]. In order to compare the outputs of both algorithms, the overlapping reads of epiG were merged using the same algorithm that is used to merge the reads in ChromaClique. The switch errors and prediction error for epiG were calculated using these merged reads as explained above.

BaseLine Chromatype. In order to assess the performance of the algorithms ChromaClique and epiG, a *BaseLine chromatype* was constructed, which was composed of only closed positions. The idea of the BaseLine is to measure the error for the simplest possible predictor. The switch error and prediction error were calculated for the BaseLine chromatype in the same way. The percentage of coverage was varied for the BaseLine chromatype to simulate insufficient coverage scenarios.

4 Results

We generated simulated data for the evaluation of the epiG and ChromaClique algorithms. First, we compared the relationship between transition rates and distances between our simulated data and real HepG2 NOMe sequencing data (Fig. 4). As expected, the probability of observing a transition goes up as the distance increases between two consecutive GCH occurrences and plateaus at a certain value. This general trend is observed for both the real and simulated data.

We then compared the performance of ChromaClique on the simulated datasets to epiG [9], which was run in "NOMe-seq" mode with the minimum number of GCH positions (`min_DGCH` flag) set to 2. The way epiG outputs chromatypes is different from ChromaClique and therefore some post-processing was

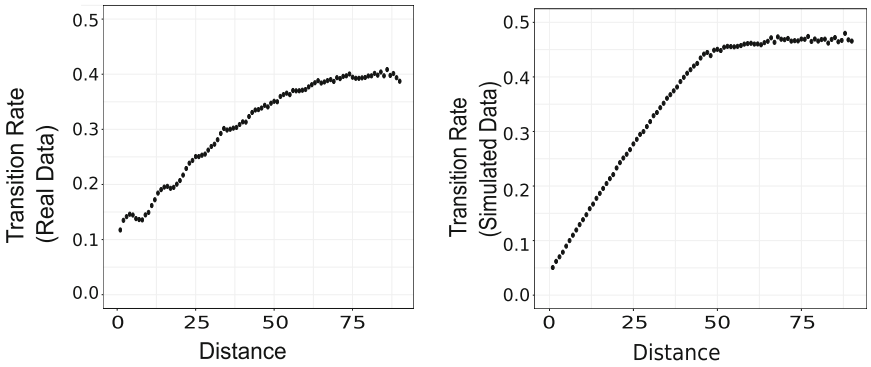

Fig. 4. Plots showing the transition rates at different distances between consecutive GCH occurrences for forward strand reads mapped to chromosome 1 for HepG2 data (left) and simulated data (right).

required to compare the two algorithms. epiG assigns each read to an epigenetic haplotype (comparable to a chromatype). All reads belonging to a particular epigenetic haplotype were merged (in overlapping regions), and this was considered as a reconstruction of a chromatype. Merging the overlapping reads was done using the same merging algorithm as in ChromaClique. Each merged read group from epiG was evaluated in the same way as each super read reported by ChromaClique. The performance of a *BaseLine chromatype* containing only closed positions over the length of the considered region was evaluated as a control for the performances of ChromaClique and epiG.

The evaluation was done using the prediction error, which denotes the average number of switch errors obtained for all predicted super reads of a method (see subsection 3.3). Another criterion for evaluation of the performance of the different algorithms is the fraction of C/O positions in the original genomic region that was covered by the reconstructed chromatypes. In this way, we can assess the trade-off between a low switch error rate and a high fraction of C/O positions covered. The threshold parameter δ in ChromaClique allows to adjust this trade-off, whereas there are no such parameters in epiG. The evaluation was restricted to a region of size 100000 bps, because epiG could not be run on the whole chromosome 1, see below.

Figure 5 shows the prediction errors of ChromaClique (green triangles) for thresholds varying from 0.000001 to 0.45, plotted against the fraction of C/O positions that were covered by the predictions. Decreasing values of δ lead to a higher fraction of GC regions being covered in the output while the errors remain constant for a certain range of thresholds. Above a certain threshold, the errors increase steadily. This behavior is noticed for all four different simulated datasets. The least prediction errors were reported for the thresholds of 0.05 and 0.07 for 100 bp and 200 bp reads, respectively.

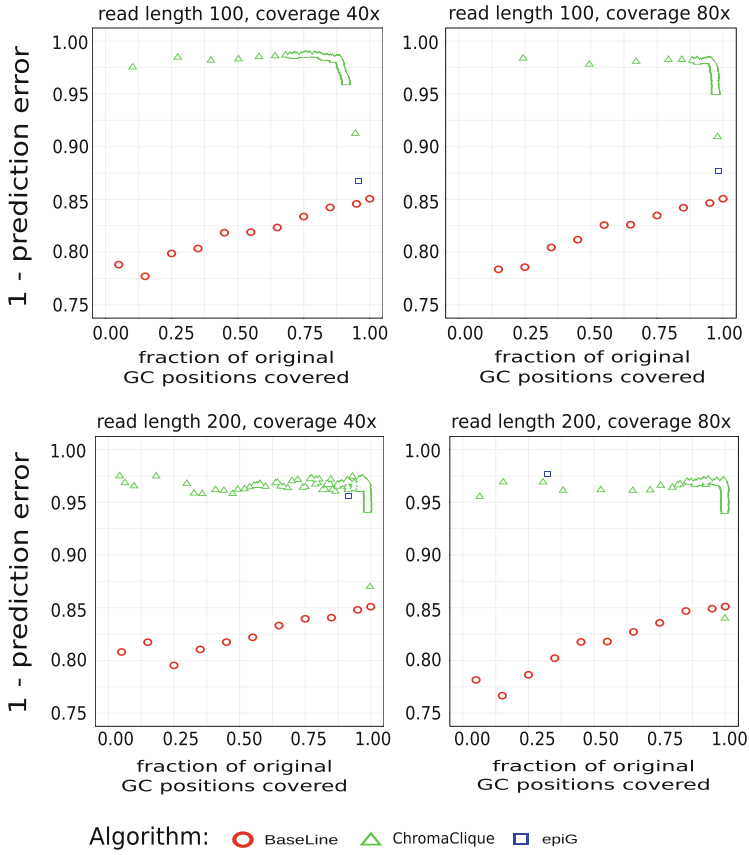

Fig. 5. Plots comparing the performance of ChromaClique with that of epiG and also a BaseLine chromatype reconstruction for four simulated data sets with different read lengths (100 or 200) and coverages (40x and 80x). (Color figure online)

We sampled varying percentages of the original GC positions to be covered by the BaseLine chromatype. In this way, we mimicked different trade-offs between error rate and fraction of covered positions, as shown by red circles in Fig. 5. For all data sets, we noticed a trend towards higher prediction error rates when fewer GC positions are covered. We observed that the number of switch errors decreases at a smaller rate than the number of GC positions covered and therefore the prediction error increases.

Figure 5 also shows the performance of epiG. Since epiG provides no parameter with which it can be tuned to get varying performances, only one error value could be obtained for each simulated dataset (blue square). For the 100 bp reads, the fraction of C/O positions covered by epiG is high at the cost of relatively high error rates, which are hardly better than the BaseLine chromatype. It seems to profit in terms of the prediction error with an increase in the length

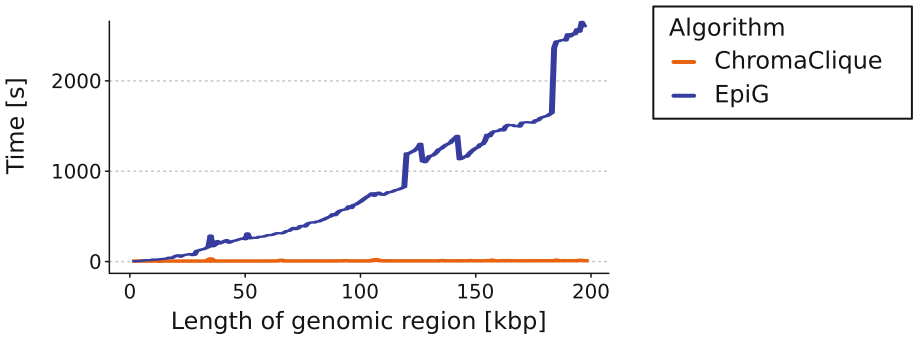

Fig. 6. Runtime of ChromaClique and epiG as a function of the length of the processed region for 100 bp reads and 40x coverage.

of the reads to 200 bps, and yields error rates which are similar to those of ChromaClique. However, an unexpected drop in C/O position coverage is noticed for the dataset with 200 bp reads and 80× coverage.

Figure 6 shows the runtimes of ChromaClique and epiG plotted against the size of the genomic region from which the initial aligned reads were sequenced. While ChromaClique's runtime grows slowly (and appears almost constant at the scale shown in Fig. 6), the runtime of epiG increases steadily with growing region sizes. While ChromaClique can be run on a chromosome-wide scale (≈101 min for the entire human chromosome 1 on 100 bp and 40X coverage data), the runtime for epiG becomes prohibitively large for regions more than 1 million base pairs.

5 Discussion and Conclusion

In this paper, we introduced ChromaClique, a novel algorithm to reconstruct nucleosome profiles from NOMe-seq data. ChromaClique is the first tool that scales to whole genomes. Furthermore, it outperforms epiG, the only competitor, in terms of prediction error rates and prediction completeness.

ChromaClique comes with the advantage that it only considers read pairs that have a sufficient C/O position overlap and then predicts whether the overlapping reads originate from the same chromatype. In contrast, epiG takes all provided reads and decides which chromatype a read is to be assigned to based on a likelihood score. That is, epiG assigns every read to a chromatype, but does not output information on where the chromatype reconstructions are reliable.

We note that NOMe-seq provides information about open and closed nucleosome positions based on the GCH regions. It hence comes with the intrinsic limitation of not being able to provide any information in GCH deserts. Thus, the reconstruction of nucleosome profiles is not possible in regions of low GC density using this protocol and we consider extending ChromaClique to accommodate other data types a fruitful direction for future research.

The runtime of ChromaClique depends on the number of cliques in the NOMe read graph, where an edge between two reads is defined by read overlaps. The number of cliques can potentially increase exponentially with an increase in the coverage. For constant coverage, however, ChromaClique scales linearly with the length of the considered region (in practice). epiG takes a different approach in its optimization algorithm. Starting from all reads as singletons initially, it optimizes for chains of reads that are overlapping each other using a likelihood formulation that uses priors on preferred lengths of read chains to search through the large space of possible combinatorial configurations. Thus, the optimization algorithm in epiG depends on the initial size of the region selected, as non-overlapping reads are considered to be part of the same haplotype chain throughout the algorithm. Our experiments suggest that for moderate to high coverage values, the speed of ChromaClique is sufficient and scales much better than the approach taken in epiG.

ChromaClique has shown a consistent performance across the different simulated datasets in terms of prediction error and the length of C/O positions covered. It consistently achieves lower error rates than epiG with the 100 bp reads. For the 200 bp reads, epiG shows similar error values to ChromaClique but lower coverage of C/O positions for the 80x case. One of the advantages of ChromaClique over epiG is its ability to tune the performance using the threshold parameter. This allows users to employ different thresholds for different datasets. For our experiments with simulated data, the thresholds that were most effective were between 0.05 to 0.07.

ChromaClique is a new method which allows for the reconstruction and subsequent analysis of nucleosome profiles on a chromosome-wide scale. In future work, it would be interesting to improve the simple simulation strategy by designing a more realistic simulation scenario, by combining real NOMe-seq data sets of different conditions. It would also be interesting to extend the model to consider DNA methylation at CpG residues as well. A promising application domain of ChromaClique is single cell NOMe-seq data, which we plan to explore in the future.

Acknowledgments. We thank Karl Nordström, Gilles Gasparoni and Jörn Walter for providing access to the HepG2 NOMe-seq data.

References

1. Thurman, R.E., et al.: The accessible chromatin landscape of the human genome. Nature **489**(7414), 75–82 (2012)
2. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J.: Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods **10**(12), 1213–1218 (2013)
3. Kelly, T.K., Liu, Y., Lay, F.D., Liang, G., Berman, B.P., Jones, P.A.: Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. **22**(12), 2497–2506 (2012)

4. Taberlay, P.C., Statham, A.L., Kelly, T.K., Clark, S.J., Jones, P.A.: Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. **24**(9), 1421–1432 (2014)
5. Durek, P., et al.: Epigenomic profiling of human $CD4^+$ T cells supports a linear differentiation model and highlights molecular regulators of memory development. Immunity **45**(5), 1148–1161 (2016)
6. Guo, H., et al.: DNA methylation and chromatin accessibility profiling of mouse and human fetal germ cells. Cell Res. **27**(2), 165–183 (2017)
7. Schmidt, F., et al.: Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. Nucleic Acids Res. **45**(1), 54–66 (2017)
8. Pott, S.: Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. eLife **6**, e23203 (2017)
9. Vincent, M., et al.: epiG: statistical inference and profiling of DNA methylation from whole-genome bisulfite sequencing data. Genome Biol. **18**(1), 38 (2017)
10. Domingo, E., Sheldon, J., Perales, C.: Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. **76**(2), 159–216 (2012)
11. Beerenwinkel, N., Günthard, H.F., Roth, V., Metzner, K.J.: Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. **3**, 329 (2012)
12. Posada-Cespedes, S., Seifert, D., Beerenwinkel, N.: Recent advances in inferring viral diversity from high-throughput sequencing data. Virus Res. **239**, 17–32 (2017)
13. Töpfer, A., Marschall, T., Bull, R.A., Luciani, F., Schönhuth, A., Beerenwinkel, N.: Viral quasispecies assembly via maximal clique enumeration. PLoS Comput. Biol. **10**(3), e1003515 (2014)
14. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)
15. Huang, W., Li, L., Myers, J.R., Marth, G.T.: ART: a next-generation sequencing read simulator. Bioinformatics **28**(4), 593–594 (2012)
16. Krueger, F., Andrews, S.R.: Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics **27**(11), 1571–1572 (2011)