# MethColor: a computational approach to uncover DNA methylation heterogeneity

Huy Q. Dinh[1,2], Ortrun Mittelsten Scheid[2] and Arndt von Haeseler[1]

[1]Center for Integrative Bioinformatics, Max F. Perutz Laboratories, Vienna, Austria
[2]Gregor Mendel Institute of Molecular Plant Biology, Vienna, Austria

**Abstract:**
DNA methylation is an important epigenetic biomarker, for instance in cancer medicine, but the degree of modification at a specific genomic cytosine may vary according to individual cells, tissues, or developmental stages. To capture and elucidate the heterogenous nature of DNA methylation is one of the challenges in both, biological and computational aspects. Deep sequencing data after bisulfite conversion of non-methylated cytosines which provide the methylation state at single-bp resolution allow to tackle this problem since individual reads represent individual genomic copies, even if the DNA was prepared from multiple cell types. We introduce a computational approach to identify the minimal number of distinct methylation profiles in such pooled data sets. We use a graph coloring method, called MethColor, together with empirical heuristics to characterize different, cell- or tissue-type-specific methylation profiles. A simulation study demonstrates the potential of this method.

## 1 Introduction

DNA methylation is an important epigenetic mark that plays a crucial role in the cellular development of many eukaryote organisms [Lai10]. It is also known as a crucial biomarker in nearly all types of cancers [JB03]. DNA methylation profiling as a diagnostic tool has become more and more attractive along with the development of simplified, accelerated and cost-efficient high-throughput data technology. DNA methylation is widely-known as the addition of a methyl group to the position 5 of at genomic cytosines. Hence, the modification leaves the DNA sequence unaltered but changes chromatin features and gene expression. DNA methylation is detected by bisulfite conversion [FMM$^+$92] which converts un-methylated cytosines (C) to uracil, after PCR to thymines (T), while the methylated cytosines are not affected. With the advent of the next-generation sequencing technology, the so-called bisulfite deep sequencing (BS-Seq) technology [LOTF$^+$08] has recently been developed to obtain high resolution DNA methylation maps. To obtain the map, the BS-Seq short reads (thereafter referred to as reads) are mapped to the reference genome. Then, the methylation profile for every genomic cytosine is generated: C-T mismatches indicate genomic unmethylated Cs whereas C-C matches indicate methylated genomic Cs in absence of sequencing error. Thus, one can characterize the methylation state for every cytosine at single-bp resolution (so-called whole-genome methylation profiles or methylomes). This is referred to as methylation profiles thereafter.

Current sequencing technologies use DNA samples prepared from a mixture of cells with potentially heterogenous DNA methylation profiles [Lai10, PE10]. This leads to average measurements across DNA molecules but inaccurate estimation of DNA methylation levels for single sites as the frequency of cell types in the mixture is typically not determined [Lai10]. This has raised a challenge in both computational and biological aspects, although each read in the BS-Seq approach provides a discrete DNA methylation pattern for a single genomic DNA molecule. Thereby, inferring the distinct cell type-specific profiles in DNA mixtures is important for uncovering the heterogeneity of DNA methylation.

Here we suggest a computational approach, called MethColor, to differentiate the methylation profiles across cells in the DNA mixture when given a mapped read library. First, the mapped read library is transformed into a graph in which each node represents one read and an edge is formed if the reads overlap on the reference genome but display a different methylation patterns (i.e incompatible reads). Then, we propose an optimization problem of finding the minimal number of distinct methylation profiles as a graph coloring problem in computer science [CLR90]. The read nodes that have the same color constitute a methylation profile. Different colors imply distinct profiles in the cell population, and the minimal number of profiles provides the most parsimonious explanation of the observed methylation state from the mapped reads. In addition, we use a simple heuristic based on the empirical methylation frequency at a single locus estimated from the mapped reads to obtain the final methylation profiles. To demonstrate the efficiency of our approach, we use simulated data and evaluate both the minimal number of inferred profiles and the similarity between the inferred profiles and the original. The results offer a promising perspective for the proposed approach in understanding DNA methylation heterogeneity.

## 2 Problem Formulation

### 2.1 Input data and Optimization problem

The input data is a reference genome consisting of $n$ genomic cytosines and a library of mapped BS-Seq reads, where upon we only consider reads that map uniquely to the reference genome without sequencing errors (the sequencing errors can be corrected [EPM$^+$08, MSKP11]). Each mapped read $r$ is represented by a so-called location vector $p(r) = (p_1(r), \ldots, p_k(r))$, $p_1(r) < \ldots < p_k(r)$ which indicates the positions of the cytosines in the reference genome where $k$ is the number of genomic cytosines mapped by read $r$. We use $p_i$ instead of $p_i(r)$ for short. Each mapped read has an associated methylation string of length $k$, $S(r) = s_1 \ldots s_k$ where

$$s_i = \begin{cases} 1 & \text{the cytosine at the genomic position } p_i \text{ is methylated} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The goal is to find the minimal set of methylation profiles $\mathcal{G} = \{G_1, \ldots, G_\gamma\}$, where $G_i = g_{i1}, \ldots g_{in}$ with $g_{ij} \in \{0, 1\}, j = 1 \ldots n$ indicates the methylation state of the cytosine at position $j$, such that the methylation string of every read is the substring of

at least one profile in $\mathcal{G}$. In other words, one needs to find a minimal number of distinct DNA methylation profiles of the whole genome, such that the methylation pattern of every read is contained in at least one profile. This problem is classified as a NP-hard problem ([PS10] and references therein).

Figure 1a shows an input of 9 reads across 5 genomic cytosines. The methylation strings derived from the corresponding reads are displayed in Figure 1b. For this example, three profiles $G_1 = 11000, G_2 = 00011, G_3 = 10101$ are the unique optimal solution to explain the data. Specifically, the reads $\{R1, R4, R7\}$ are generated by $G_1$, $\{R2, R5, R8\}$ by $G2$ and the rest by $G_3$.

### 2.2 Graph coloring problem

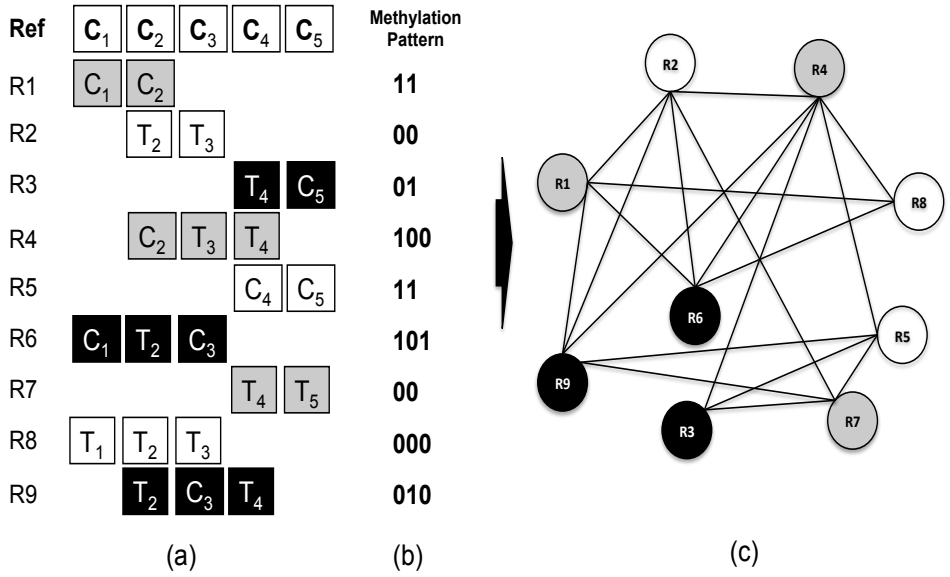



Figure 1: An illustrative example where the reference genome has 5 cytosines ($n = 5$): (a) 9 reads derived from 3 cells with different DNA methylation patterns indicated by 3 colors: grey ($G_1$), white ($G_2$), and black ($G_3$); the subscripts indicate the mapped positions at cytosines enumerated by the order of their relative positions in the reference genome; (b) the corresponding methylation pattern of 9 reads; (c) the graph with 9 nodes built from the input data, three colors are assigned to the reads according to their cell of origins.

We reformulate the above optimization problem as a graph coloring problem. Each read is represented as a node, two nodes are connected by an edge if the mapped regions in the reference genome corresponding to the two reads $r$ and $r'$ overlap. It means $\exists(p_i, \ldots, p_j)$ such that $min(p(r), p(r')) \leq p_i \leq p_j \leq max(p(r), p(r'))$ and if the two methylation patterns in the overlap are different. That is, an edge indicates that the methylation patterns

of the two reads are different although both reads map at least partially to the same genomic region. We solve the minimal number of methylation profile problem by assigning colors to the nodes such that any two nodes connected by an edge must have different colors. Figure 1c illustrates the graph for the data in Figure 1a. The nodes are colored by 3 colors corresponding to the 3 distinct profiles $G_1, G_2, G_3$ mentioned above.

# 3 MethColor method

## 3.1 Graph coloring to estimate the minimal number of methylation profiles

We suggest a greedy algorithm as a quick approximation to find the minimal number of colors to color the nodes of the read graph. The greedy idea is similar to the McColor algorithm presented in [LM05]. However, we apply it to arbitrary graphs and use a heap data structure for efficiency in choosing the coloring nodes.

Given $M$ nodes, $\mathcal{C} = \{1, \ldots, M\}$ is the set containing the maximal number of colors. For each node $x$, $nbCol(x)$ denotes the set containing the colors of its colored neighbors and $color(x)$ denotes the color of a node $x$, $color(x) = 0$ if $x$ is not yet colored. Our algorithm (see pseudo code: Algorithm 1) works as follows: we start with color 1 assigned to a random node. At each step, we select from the set of yet uncolored nodes ($color(x) = 0$) the node $x$ for which $|nbCol(x)|$ is maximal; and the color assigned to x is $color(x) = min\{\mathcal{C} \setminus nbCol(x)\}$. For every uncolored $y$ connected by an edge with $x$, $nbCol(y)$ is then updated. The process is repeated until all nodes are colored. For efficient implementation, we use the heap data structure [CLR90] as a binary tree where each node $x$ of the heap $H$ has the key value $key(x) = |nbCol(x)|$. Thus, the root of the heap is always the uncolored node $x$ with the largest $|nbCol(x)|$.

**Algorithm 1**: Pseudo code of graph coloring algorithm

**Data**: A mapped read library.
**Output**: An assignment of colors for every node.

**begin**
  Building graph $G = (V, E)$ corresponding to the mapped read library;
  Initialization: $\forall x : nbCol(x) = \emptyset$; $color(x) = 0$; $push(x, H)$;
  **foreach** $i$ = 1..$|V|$ **do**
    $x = pop(H)$ ;
    $color(x) = min\{\mathcal{C} \setminus nbCol(x)\}$;
    **forall** $y : (x, y) \in E$ *and* $color(y) = 0$ **do**
      $nbCol(y) = nbCol(y) \cup color(x)$ ;
      update $H$: $key(y) = |nbCol(y)|$ ;
**end**

The graph coloring algorithm has the worst complexity of $O(|E| \log |V|)$ where $|V|$ is the number of nodes, i.e the number of reads and $|E|$ is the number of edges. Generating the

graph has $O(|E| * k_{max})$ complexity where $k_{max}$ is length of the longest location vector.

### 3.2 Heuristics for determining the DNA methylation profiles

After coloring the read graph, the reads that belong to one color are used to construct the intermediate profile as the consensus string of those reads' methylation strings. $\hat{\mathcal{G}} = \{\hat{G}_1, \ldots, \hat{G}_\gamma\}$ is the set of inferred methylation profiles, in which $\hat{G}_i = \hat{g}_{i1} \ldots \hat{g}_{iN}, \hat{g}_{ij} \in \{0, 1, 2\}$, where 0 indicates un-methylated, 1 methylated and 2, unresolved, i.e there is no read assigned at genomic cytosine $j$ and profile $i$. We note that due to the greedy strategy, some methylation profiles are only comprising very few reads, i.e the profile consists of many 2 (unresolved states).

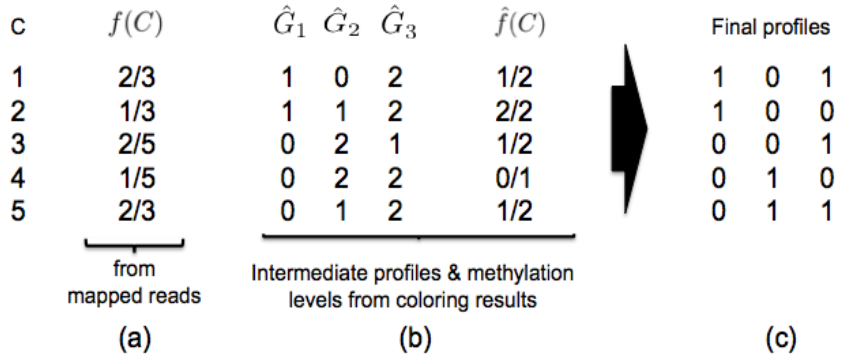



Figure 2: Illustration of building the final methylation profiles: (a) the 5 cytosines in the example from Figure 1 and their empirical methylation level $f(C)$; (b) the columns represent 3 intermediate patterns with 1: methylated, 0: un-methylated, 2-unresolved and their methylation level $\hat{f}(C)$; (c) The heuristics move reads from the $\hat{G}_1$ to $\hat{G}_2$ or $\hat{G}_3$ to resolve unresolved positions such that $\hat{f}(C_i)$ is as close to $f(C_i)$ as possilbe. This leads to the final profile.

To determine unresolved methylation states, we use a heuristic based on the empirical methylation level $f(C_i)$ at cytosine position $i$ that can be computed from mapped reads as the ratio of reads with cytosine at that position over the total number of mapped reads at position $i$. From the intermediate profiles we compute $\hat{f}(C_i)$ as the preliminary methylation level computed from the profiles ignoring unresolved states (Figure 2 shows an example). To minimize the difference between $f(C_i)$ and $\hat{f}(C_i)$ we apply the following heuristic. A read $r$ is called movable from one profile $\hat{G}$ to another $\hat{G}'$ if its methylation pattern is compatible with the methylation profile of $\hat{G}'$. If we move a read from $\hat{G}$ to $\hat{G}'$, we will recalculate the $\hat{f}(C_i)$ for all cytosines affected by this move. Then, the goal is now to replace most of the unresolved states in the intermediate profiles and at the same time to minimize the difference $i$ for which $|f(C_i) - \hat{f}(C_i)|$. To do this we perform a greedy strategy by starting with genomic position $i$ if the $|f(C_i) - \hat{f}(C_i)|$ are maximal. This process is repeated until we cannot find any suitable movement. Fig (2c) displays the final result.

# 4 Simulation Study

## 4.1 Datasets

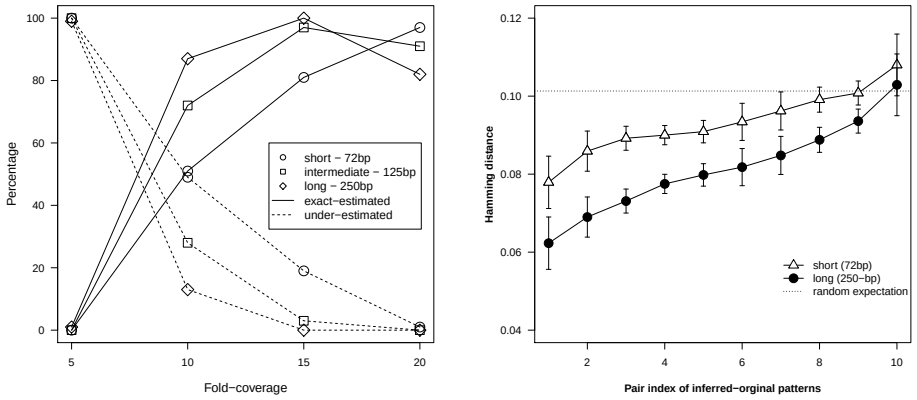
We simulated the BS-Seq short reads from 10kbp reference assuming no sequencing error, a $6\%$ methylation rate (motivated from empirical study [LOTF$^+$08]) and $10$ distinct DNA methylation profiles. Mapped short (72-bp), intermediate (125-bp), and long (250-bp) reads were randomly generated with sequencing coverage of $5, 10, 15$, and $20$ after mapping. A coverage of $20$ is high compared to the available BS-Seq datasets [LOTF$^+$08, LPK$^+$11]. Here we simulated equal number of reads for each profile. Hundred datasets were generated for each combination of read length and sequencing fold-coverage. Thus, we had $1200$ datasets in total.

For evaluation, we first computed the empirical distribution of number of inferred methylation profiles for all simulations, i.e how often MethColor estimated exactly the original number of methylation profiles or under-/overestimated it. Then, we used Hamming distance to evaluate the similarity between inferred and originally simulated methylation profiles. As the assignment of the inferred profile and the original profile is not predictable, we applied a greedy strategy. First, the pair of predicted and simulated profiles with smallest Hamming distance was chosen, deleted from the set of unassigned ones and then the process was repeated until no pair can be selected. Then we tested if the Hamming distance for each selected pair is significantly smaller in comparison with the distribution of Hamming distances between random patterns at the same methylation rate.

## 4.2 Simulation results

Figure 3(a) shows that the ability to predict the correct number of profiles depends on the read length and the coverage. In case of low coverage (of coverage $5$), all the test underestimate the number of distinct profiles regardless the read length. Increasing number of mapped reads leads to more cases in which the algorithm estimates exactly the number of profiles. In addition, longer reads also improve the coloring results for the case of high coverage (sequencing fold of $10$ and $15$). However, when the coverage exceeds 15 the precision of our algorithm decreases, it tends to overestimate the number of profiles especially for long read lengths. We speculate that the graph gets to complex due to too many reads and too many edges. This simulation also provide a hint of how much sequencing is needed to have an exact estimation of methylation profiles.

Finally, not only the precise number of different profiles is estimated but also the accuracy of individual methylation state of each profile. Figure 3(b) presents the average Hamming distance of $10$ inferred-original pairs indexing from the closest pair to the most distant one. Here we only show two examples from simulated datasets (the short and the long read case with the sequencing coverage of $20$), that provide the best fit to the simulated profiles (small Hamming distance, lower curve) and the worst fit (large Hamming distance, upper curve). The results demonstrate that MethColor can estimate well the methylation




(a) Number of inferred methylation profiles (b) Similarity between inferred and original profiles

Figure 3: Performance of MethColor with diverse sequencing coverages and read lengths: (a) solid/dash lines indicates number of exactly/under-estimated number of cell-types, (b) The average Hamming distance of the inferred patterns and original ones for the worst and the best cases that corresponds to the short and long read length at 20-fold coverage. The horizontal dashed-line indicates the $5\%$ lower bound of Hamming distance between two random profiles.

patterns of up to $8$ out of $10$ cell-types as the Hamming distance is ranging from $0.06$ to $0.1$. The average Hamming distance is then significantly smaller than the analytical lower bound from the distribution of distance between random patterns given the same methylation rate (dotted horizontal line in figure 3(b)), assuming a p-value $0.05$.

We also did the experiments with non-uniform distribution of cell-type frequency, and the results are similar. Taking all together, the simulation results prove the potential MethColor to be applied for experimental biological data.

## 5 Discussion

We formulated a computational problem aiming to understand the heterogeneity of DNA methylation based on the mapping profiles from deep sequencing data after bisulfite conversion. Our approach, MethColor, efficiently estimates the number of distinct cell-types as well as their methylation profiles. Despite a rough estimation, this can provide valuable information for research and diagnosis and help to understand the diversity of DNA methylation patterns.

In addition, the cell-type frequency can be estimated based on the inferred patterns by the Expectation Maximization approach used in haplotype construction problem [EPM$^+$08]. Due to the page limitation, we do not show the results here. Our approach can also be generalized to the haplotype reconstruction, for instance, in the viral population according

to [EPM+08]. Our approach can also work in context of pooled sequencing of different individual or ecotypes, in presence of one available reference genome, e.g. help to assemble short-reads mapped to common reference genome [PS10].

Future works will address: (i) incorporating sequencing errors, (ii) applying the MethColor to analyze available biological datasets, for example the recent diverse set of methylome data from human embryonic stem cells [LPK+11]. In dealing with sequencing error, we can either do error correction before or incorporate the non C-T mismatches as the weights of the read graph.

**Acknowledgment.** This work was supported by a grant from the Wiener Wissenschafts-, Forschungs- and Technologiefonds (WWTF) to A.v.H. The PhD grant to H.Q.D is also financed by the GMI to O.M.S. We are grateful to Anh Nguyen and Minh Bui for useful discussions on the manuscript.

## References

[CLR90] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to algorithms*. The MIT Press/McGraw-Hill, Cambridge, Masachusetts, 1990.

[EPM+08] N. Eriksson, L. Pachter, Y. Mitsuya, S. Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel. Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, 4:e1000074, Apr 2008.

[FMM+92] M Frommer, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, and C L Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA*, 89(5):1827–31, Mar 1992.

[JB03] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, 33 Suppl:245–254, Mar 2003.

[Lai10] P. W. Laird. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, 11:191–203, Mar 2010.

[LM05] Benjamin Lévêque and Frédéric Maffray. Coloring Meyniel graphs in linear time. *Electronic Notes in Discrete Mathematics*, 22:25 – 28, 2005. 7th International Colloquium on Graph Theory.

[LOTF+08] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–36, May 2008.

[LPK+11] R. Lister, M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M. Evans, and J. R. Ecker. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471:68–73, Mar 2011.

[MSKP11] P. Medvedev, E. Scott, B. Kakaradov, and P. Pevzner. Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics*, 27:i137–i141, Jul 2011.

[PE10] Mattia Pelizzola and Joseph R. Ecker. The DNA methylome. *FEBS Letters*, In Press, Corrected Proof:–, 2010.

[PS10] Qian Peng and Andrew D. Smith. Multiple Sequence Assembly from Reads Alignable to a Common Reference Genome. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 99(RapidPosts), 2010.