Supplementary Information for methylFlow: cell-specific methylation pattern reconstruction from high-throughput bisulfite-converted DNA sequencing

Faezeh Dorri Lee Mendelowitz Héctor Corrada Bravo

University of Maryland College Park

Supplementary Information

Simulation Setup

Simulating patterns

Methylation patterns are simulated such that nearby CpG sites are likely to have similar methylation status based on their genomic distance. Specifically, two CpG sites with distance d have the same methylation status with probability:

$$f(d) = 1 - \frac{1}{1 + e^{-10 \times (d - \mathrm{corrDist})}} \tag{1}$$

where $\mathrm{corrDist} = 20$ is a parameter that controls methylation status correlation between two consecutive CpG sites. % It varies between different simulation settings, i.e. for simple setting, since we want two very distinct patterns the $corrDist$ is larger than other settings. To simulate a pattern, the methylation status of the first CpG is set uniformly at random, and each subsequent site is set to the same status as the previous CpG with probability $f(d)$, othwerwise it is set uniformly at random. CpG locations for 1750 CpGs were obtained from a whole genome bisulfite sequencing dataset [10].

Simulating short reads

For simulating the sequencing process, we randomly select a pattern with probability proportional to its abundance. Read start position is uniformly chosen at random. Every CpG site is sequenced without any error with probability $1 - \mathrm{error}$. If parameter $\mathrm{error} = 0$, then methylation pattern of every short read is exactly the same as its true pattern. A total of $\frac{\mathrm{coverage} \times \mathrm{dnaLength}}{\mathrm{readLength}}$ short reads are generated in each simulation setting.

Parameters for coverage, short read length, number of CpG sites in the simulation region are varied over a specified range for each simulation setting as we test the behavior of the algorithm

as a function of these three parameters. Otherwise, these parameters are held to constant values 20 for coverage, 100 for number of CpGs, and 70 bp for short read length.

When testing the effect of each of these parameters on the performance of our algorithm, coverage is varied from 5 to 20, short read length varies from 50 to 250, the number of CpG sites varies from 75 to 150.

Minimum cost network flow error metric

Our third error metric evaluates performance based on both methylation call error and pattern abundance estimates. In this metric there is no threshold is used to filter pattern matches between simulated and estimated patterns. Instead, we run a minimum cost network flow problem that matches every true pattern to a set of estimated patterns on the same bipartite graph.

However, we add constraints such that the sum of outgoing flows from every node in true pattern set is equal to the abundance of corresponding true pattern and the sum of incoming flow to every node of estimated pattern is also equal to the abundance of corresponding estimated pattern. f_{ij} is the amount of flow goes from true pattern i to pattern j such that the sum of all f_{ij} s are minimized. Then we compute the expected methylation call error by multiplying the probability of pattern i and j being matched, i.e., what percentage of the abundance of pattern i is covered by pattern j. The cost of our network flow in our bipartite graph is equal to sum of the weights of every pair (i, j) multiplied by the amount of the corresponding flow. This metric evaluates how well our algorithms predict both methylation calls and the abundances.

$$\begin{aligned} \text{cost of network} &= min \sum_{ij} w_{ij}.f_{ij} \\ \text{s.t} \sum_{j} f_{ij} &= \theta_i, \forall i \in S \\ \sum_{i} f_{ij} &= \theta_j, \forall j \in T \\ f_{ij} &\geq 0, \forall i \in S, \forall j \in T \end{aligned}$$

Here f_{ij} is the amount of flow from node $i \in S$ to node $j \in T$ and corresponds to the fraction of the abundance of simulated pattern i matched to estimated pattern j. The cost of our network flow in our bipartite graph is equal to sum of the weights of every pair (i, j) multiplied by the amount of the corresponding flow. It is a measurement to evaluate how well our algorithms predict both methylation calls and the abundances.

Supplementary Figures

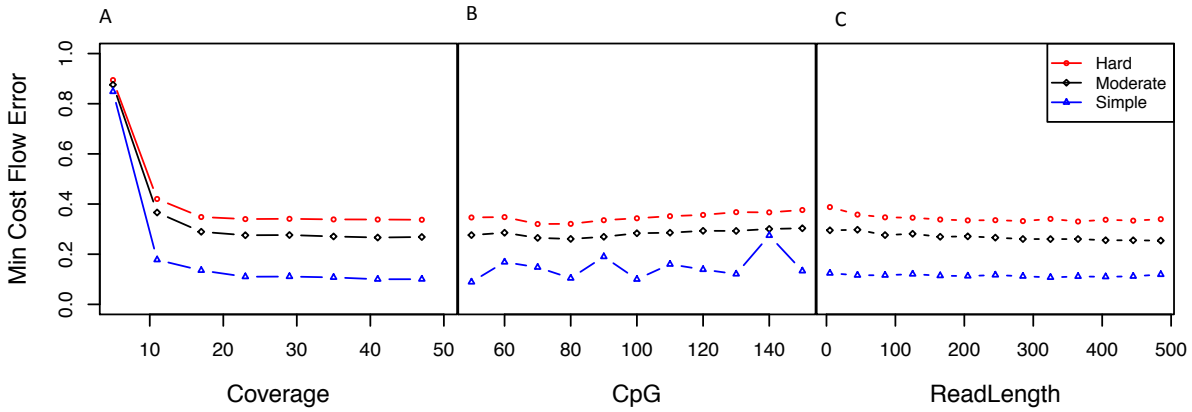

Figure 1: Minimum cost flow (MCF) error for three different simulation settings with different complexity. (A) The effect of coverage on MCF error. (B) The effect of the number of CpG sites on MCF error. (C) The effect of short read length on MCF error.

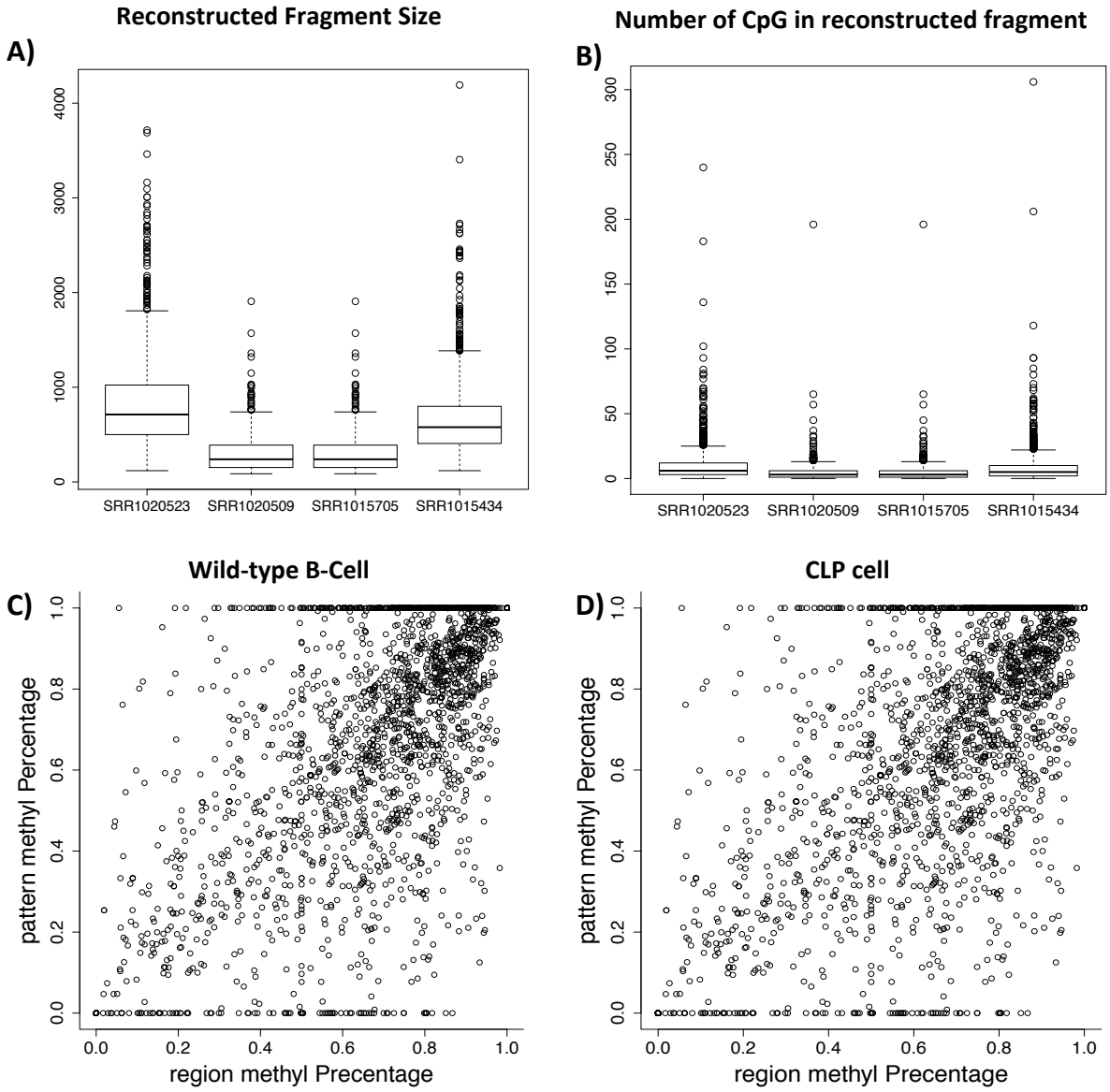

Figure 2: Pattern estimation in whole genome bisulfite sequencing of wild B-cells and KSL/CLP cells. (A) Length distributions of reconstructed cell-specific methylation patterns from chromosome 3. (B) Distributions of the number of CpGs per reconstructed cell-specific methylation patterns from chromosome 3. (C and D) CpG methylation percentage estimated from reconstructed cell-specific methylation patterns *(pattern methyl Percentage)* vs. observed CpG methylation percentage *(region methyl Percentage)* for wild type sample sample and KSL cells, respectively.