Analysis of methylation data with Imputation with blocks (MethyBlock).

Wei Meng#, Mithlesh Ray#, Christopher Fenton, Endre Anderssen, Ruth Paulssen*

Clinical Bioinformatics Research Group, Department of Clinical Medicine, Faculty of Health Sciences, UiT- The Arctic University of Norway, Tromsø, Norway.

Background

Analysis of methylation data is dependent on two inputs. The total number of reads at any given cytosine site (coverage), and how many of those cytosines are methylated (methylation). Both the large number of sites and low coverage can cause problems in the analysis of methylation data. Likewise finding DMR is computationally and statistically challenging. Here we present a new method (MethyBlock) that highlights candidate regions by balancing co-variation and chromosomal distance.

Methods

Raw data were processed by Bismark. Sample sites under a user-defined coverage were set to zero. Furthermore, sites with too many coverage zeros across all samples were removed. Data is divided into chromosomes, each chromosome divided into segments based on the distance to the neighboring CpG site. Relative methylation level data is calculated by dividing methylation counts by coverage counts. Missing values were imputed using K-nearest neighbor (KNN) per segment. Imputed segments were divided into blocks by balancing co-variability and distance. Blocks were further filtered for outliers. The blocks and the imputed relative methylation matrix are kept for further analysis.

Results

With imputation on a certain percentage of the samples, MethyBlock returned more base pairs but grouped into smaller regions than the DMRs generated from DMRseq. This may allow the finding of more specific blocks within large CpG regions. MethyBlock also returned a similar percentage of CpG islands, regulatory regions, and functional annotations. Comparison with DMRseq shows a similar percentage in UCSC hg38 overlapped regulatory regions.

Conclusions

MethyBlock simplifies the downstream computational and statistical analysis by reducing the data to the region level. This method improves the power of statistical tests by reducing the impact of multiple testing.

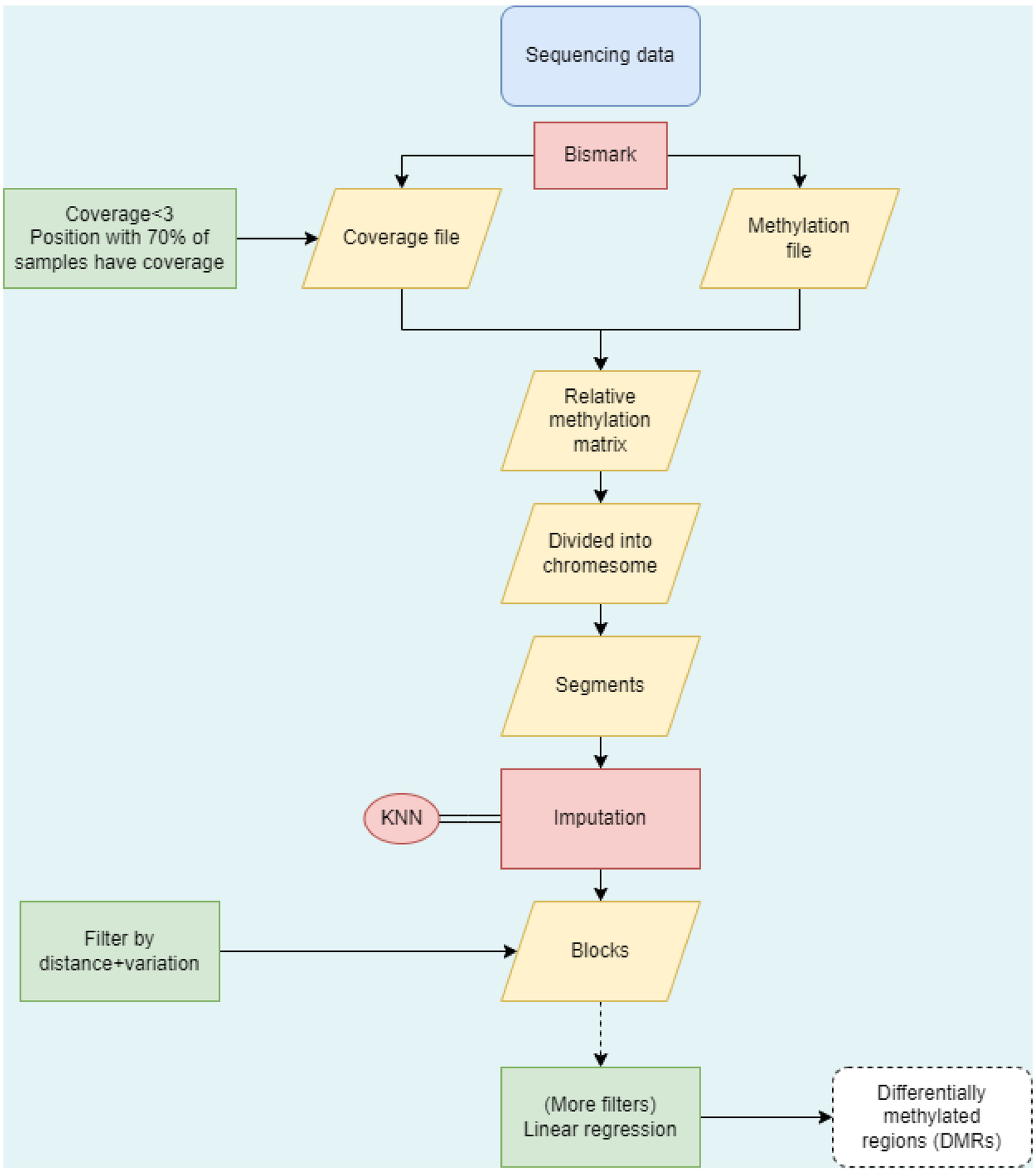

Table 1.
Stats of the blocks. Total regions are the number of clustered methylation sites counted as regions. Region width is the average length of the sites. Base pairs are the number of the sites included. Average CpGs are the percentage overlapping with CpG islands. The imputation percentage stands for how much of the data was imputed.

	90%	80%	70%	60%	50%	DMRseq
Total regions	37944	48121	55566	62159	67334	15925
Region width	170.6	182.7	194.8	207.8	223.0	533.3
Base pairs	6472760	8794022	10822046	12917813	15015613	8493345
Average CpGs	19.4	21.5	23.3	25.2	27.2	23.6
Imputation %	0.8	2.6	5.2	7.9	10.6	-

Table 2. Overlapping areas of the blocks. The UCSC regulatory regions were used as a reference to examine how many blocks are responsible for regulation. Last column stands for the overlapping percentages of DMRs found by dmrseq.

	90%	80%	70%	60%	50%	DMRseq
Open chromatin region %	1.51	1.51	1.57	1.62	1.65	1.76
TF binding site %	1.72	1.75	1.78	1.79	1.83	1.28
CTCF binding site %	14.33	14.68	14.97	15.09	15.31	18.04
Enhancer %	2.85	2.75	2.70	2.76	2.82	4.82
Promoter %	31.70	29.75	28.22	27.07	25.99	28.20
Promoter flanking region %	13.53	13.58	13.74	13.84	13.87	22.29

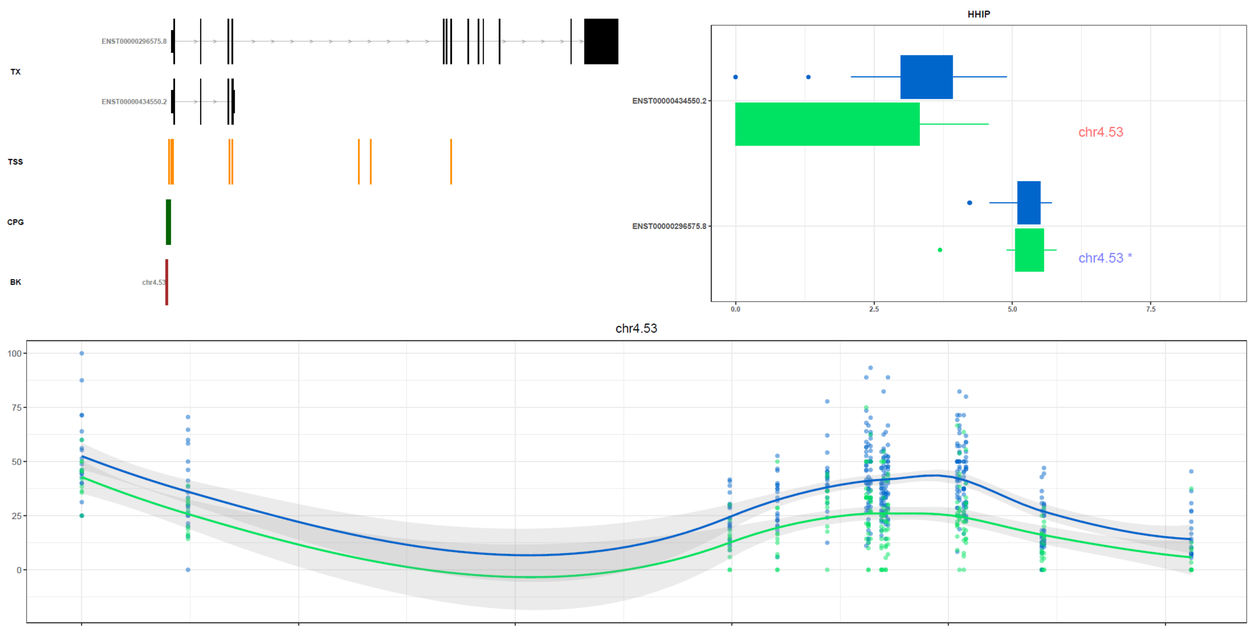

Figure 1. Possible outcome with visualization. The figure depicts the correlation between methylation status from blocks and the transcriptomic data. The transcript and regulatory position of the transcript is aligned with identified blocks on the top left, showing the regional transcript information. The data type is listed on the left: TX is demonstrated as the transcript position. Larger bars indicate introns in TX. TSS stands for the transcript starting site. CPG stands for the CpG island in a transcript. BK is the region found named with chromosome and a tag. The differential expression level is shown in the top right with the bar's regions. Different transcripts of one gene (as the name of the title) are indicated if multiple transcripts are involved. The region name is marked with a correlation coefficient from -1 to 1, blue to red, respectively. The X-axis shows the log normalised expression level. The differential methylation level of each region is shown at the bottom. The X-axis is the position of methylated sites in the region, and Y-axis is the methylation percentage. Each dot is one percentage of methylation position in each sample. The local regression is shown as a smoothed line with blue and green as groups. The grey area of the line stands for 95% confidence level interval for predictions from the regression.