seqlm: an MDL based method for identifying differentially methylated regions in high density methylation arrays

Kaspar Märtens*[a], Raivo Kolde*[ac], Kaie Lokk [b], Sven Laur [a], Jaak Vilo [ac]
[a] Institute of Computer Science, University of Tartu, Estonia
[b] Institute of Molecular and Cell Biology, University of Tartu, Estonia
[c] Quretec Ltd
* These authors contributed equally
Contact: raivo.kolde@eesti.ee, kaspar.martens@gmail.com

Euroopa Liit Euroopa Regionaalarengu Fond
Eesti tuleviku heaks
DoRa
ARCHIMEDES

1. Introduction

One of the main goals of large scale methylation studies is to detect differentially methylated loci. In the novel methylation arrays such as Illumina 450K the density of probes is high enough that it is possible to identify differentially methylated regions (DMRs) containing multiple probes from the array.

Compared to the standard approach, where the differential methylation is tested on each probe separately, the task is more complex. In addition to measuring differential methylation one has to combine the probes into reasonable regions. Few tools exist for DMR identification from this type of data, but there is no standard approach.

We propose a novel method for DMR identification. By fitting linear models within sliding windows of different lengths and choosing the best configuration according to the minimum description length (MDL) principle, we divide the genome into regionswith similar methylation patterns. The resulting regions are scored using a mixed model that takes into account correctly the repeated nature of measurements. Our approach has several advantages. Most importantly, it depends on very few parameters, thus it requires less tuning.

The method is implemented as an R package and is available in Github https://github.com/raivokolde/seqlm.

2. Why regions?

The usual approach to find differentially methylated loci is sitewise, *i.e.* a statistical test is applied to each position separately. However, searching for longer regions is for several reasons more natural.

- Methylation is regulated in longer regions (Lienert *et.al.* 2011)
- Correlated structure of methylation is apparent in the data (Figure 1A)
- Intermediate CpG sites (not measured by array) follow the same pattern that was identified in the array. (Figure 1B)
- Statistical tests have more power to detect differences in longer regions.
- The number of results is much smaller, but more informative.
- The patterns found in longer regions should be biologically more reliable.

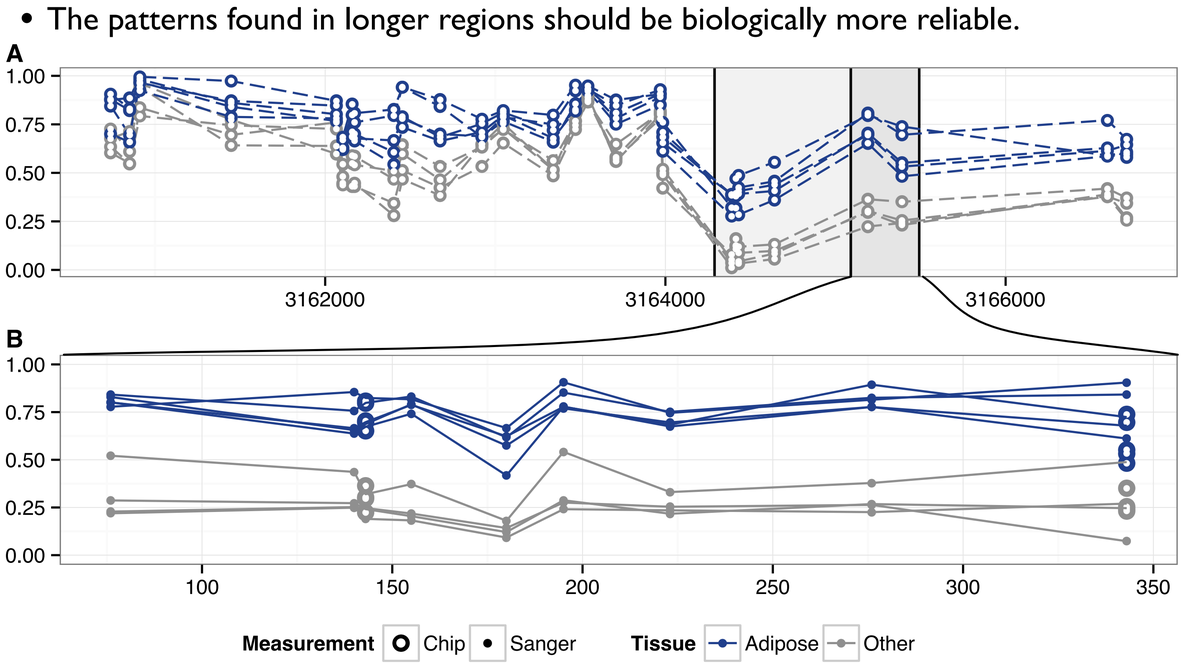

Figure 1: *Example of a genomic region, (A) shows the array measurements, with grey box indicating a DMR. (B) shows validation results using Sanger sequencing in a smaller part of the DMR.*

3. Minimum Description Length principle

Our approach is based on the Minimum Description Length (MDL) principle. This is an information theoretic criterion for model selection. It is a mathematical manifestation of *Occam's razor*. We try to find the simplest model that can accurately describe the data. To do so, we estimate how concisely the data can be described using a certain probabilistic model. The description length of a model can be expressed as a sum:

Description Length = Goodness of Fit + Model Complexity

where "*Goodness of Fit*" is given as negative log likelihood and "*Model Complexity*" is proportional to the number of parameters in the model. By minimizing the description length, we balance the accuracy of the model and its complexity.

4. seqlm method

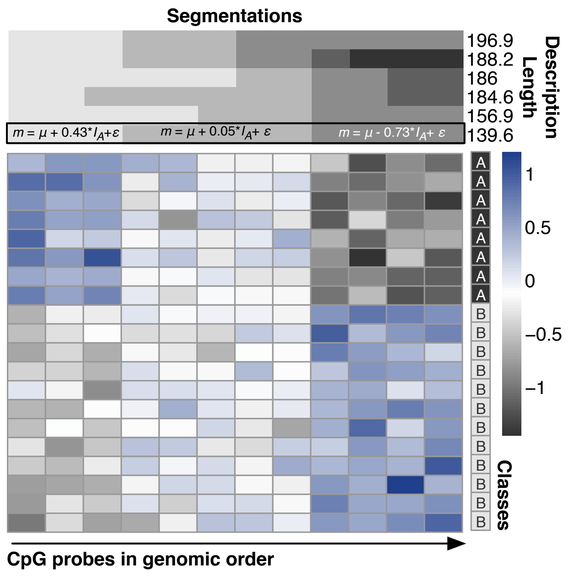

Figure 1: *Example of seqlm method of segmentation*

The seqlm method works in three stages.

Stage 1: The genome is divided into smaller pieces based on a genomic distance cutoff.

Stage 2: In each piece probes are segmented into regions that have approximately constant difference between the groups of interest.

- In sliding windows with variable sizes we fit a linear models to the data.
- For each model we record the description length - the amount of bits needed to describe the data using the model
- Using dynamic programming we find the segmentation that minimizes total description length

Stage 3: We assess the relevance of each segment, by using a mixed model where the classes are a fixed effect and a sample is a random effect. This model takes into account the repeated nature of the consecutive methylation measurements. The segments are ordered by their significance.

5. Comparison to other methods

There are two R packages available in Bioconductor, that tackle the same problem: `bumphunter` (Jaffe *et.al.* 2012) and `MethyAnalysis`. Both take a similar approach to region finding. Basically they smooth the methylation values using a sliding window, test significance per position and combine significant results together.

Compared to `bumphunter` and `MethyAnalysis`, `seqlm` has some advantages:

- it has very few input parameters, whereas results of other methods depend strongly on the smoothing window size and effect size cutoff (see Figure 3);
- it features less *ad hoc* statistical methodology for ranking the regions and it does not need to perform permutations like `bumphunter`.

Another package `IMA` tests the differential methylation significance of predetermined regions, such as gene promoters, CpG islands etc.

Drawbacks of `IMA` (Wang *et. al.* 2012) compared to `seqlm` are:

- borders of DMRs usually do not coincide with genomic annotations, therefore, `IMA` can create multiple significant regions out of one DMR (see Figure 3);
- different genomic annotations can coincide, creating overlapping regions for `IMA` (see Figure 3), that complicates the interpretation even further.

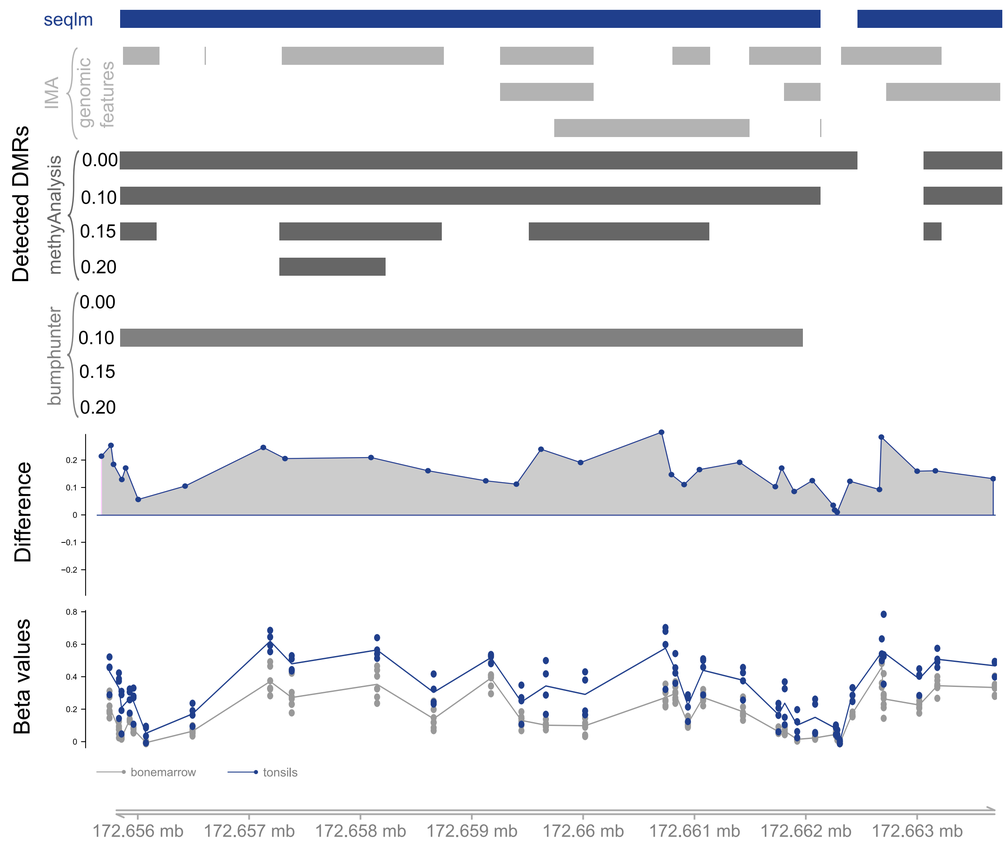

Figure 3: *An example region, showing the performance of different algorithms*

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D "*Identification of genetic elements that autonomously determine DNA methylation states.*" (2011) Nature Genetics
Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. "*Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.* "(2012) Int J Epidemiol.
Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu S. "*IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data.*" (2012) Bioinformatics