Identifying DMRs

Jonathan Michael Foonlan Tsang
Queens' College, Cambridge

November 12, 2014

The SLCU

Figure: The Sainsbury Laboratory at night. (Photo: Wikipedia, User:Cmglee)

Paszkowski Group

DNA methylation is considered as an important epigenetic mark in plants and animals. In plant Arabidopsis thaliana, a mutation in the MET1gene, encoding the main DNA methyltransferase, leads to a wide loss of DNA methylation. After re-introduction of the wild type version of MET1 gene into the mutant, DNA methylation can be only partially restored and there are chromosomal loci at which the loss of DNA methylation persists and these loci become hypomethylated epialleles that are stably inherited over many generations. We believe that the differences between loci that are able of re-methylation and loci that cannot be remethylated are encoded in their DNA sequence.

Figure: *Arabidopsis thaliana*. (Photo: Gordon Simpson, James Hutton Institute)

My project

I spent most of my time at the SLCU building a system to find DMRs between epigenomes. The system was to take in raw methylation data from the lab, process it and match it up against the (known) genome of *A. thaliana*, find the methylation profile of the specimen, and detect regions where two specimens' methylation profiles differ significantly. It would then see whether these regions matched up with regions of interest on the genome. There already exist ways procedures to do all of the above, but the SLCU had no standard procedure for going from raw data to report: different parts had been done by different people, possibly at different institutions.
I tested my system on the data that the Paszkowski group had obtained from Tuebingen.

Data

I had:

- *Arabidopsis thaliana* genome and gene annotation from TAIR
- Reads from Tuebingen (in FASTQ format)
  - Chr1-5 only (no chloroplast or mitochondria)
  - passed through Trimmomatic
  - poor quality even after trimming.
- Jacobsen's reads (also in FASTQ format, from the NCBI website)

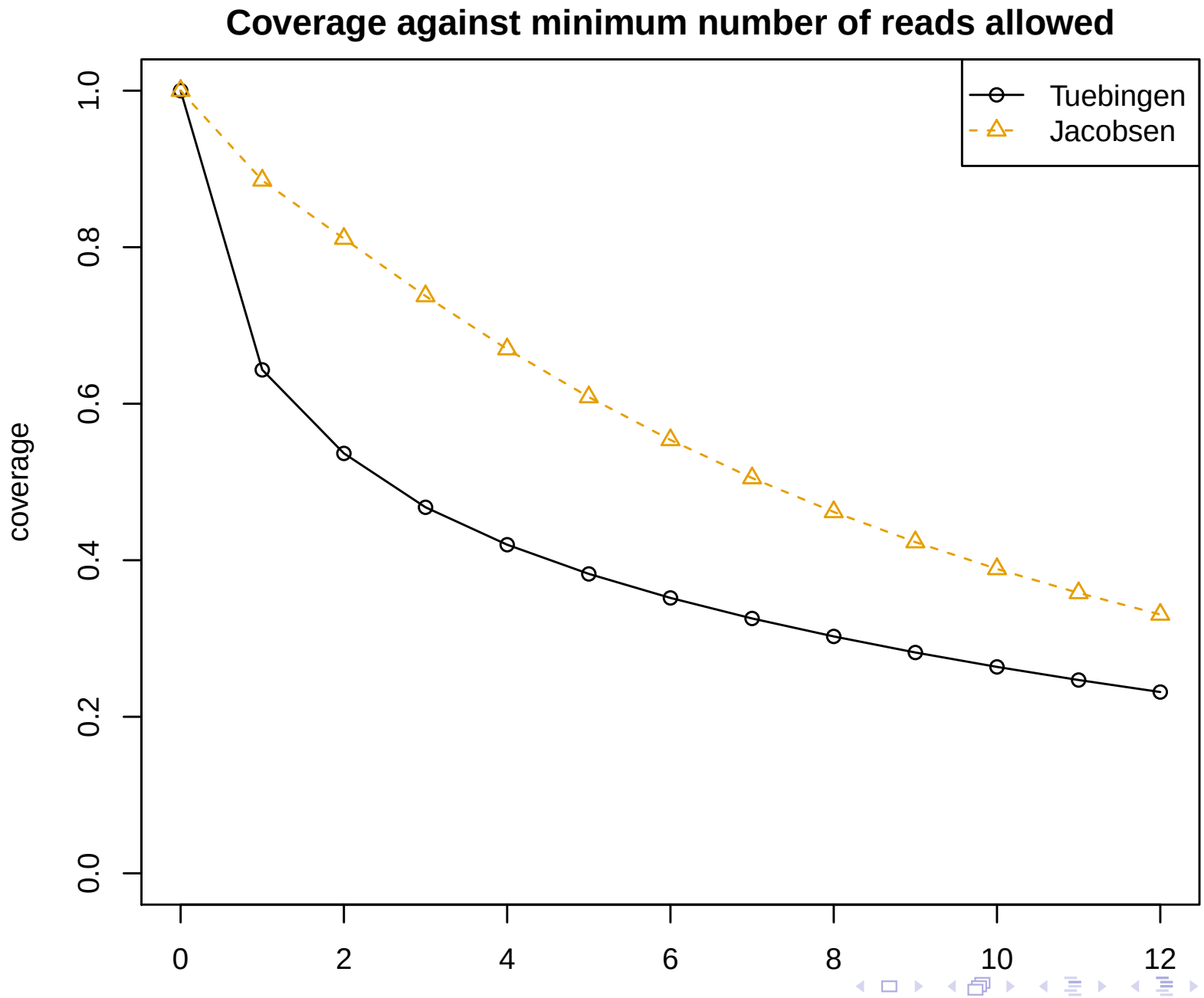
These were passed through Bismark (using Bowtie2 with -N 1 -L 28 -X 1000), to align the reads and obtain a methylation report, with the positions of cytosines, their context, the number of reads on each position in each condition, and the number of those which show methylation.

Problems with the data

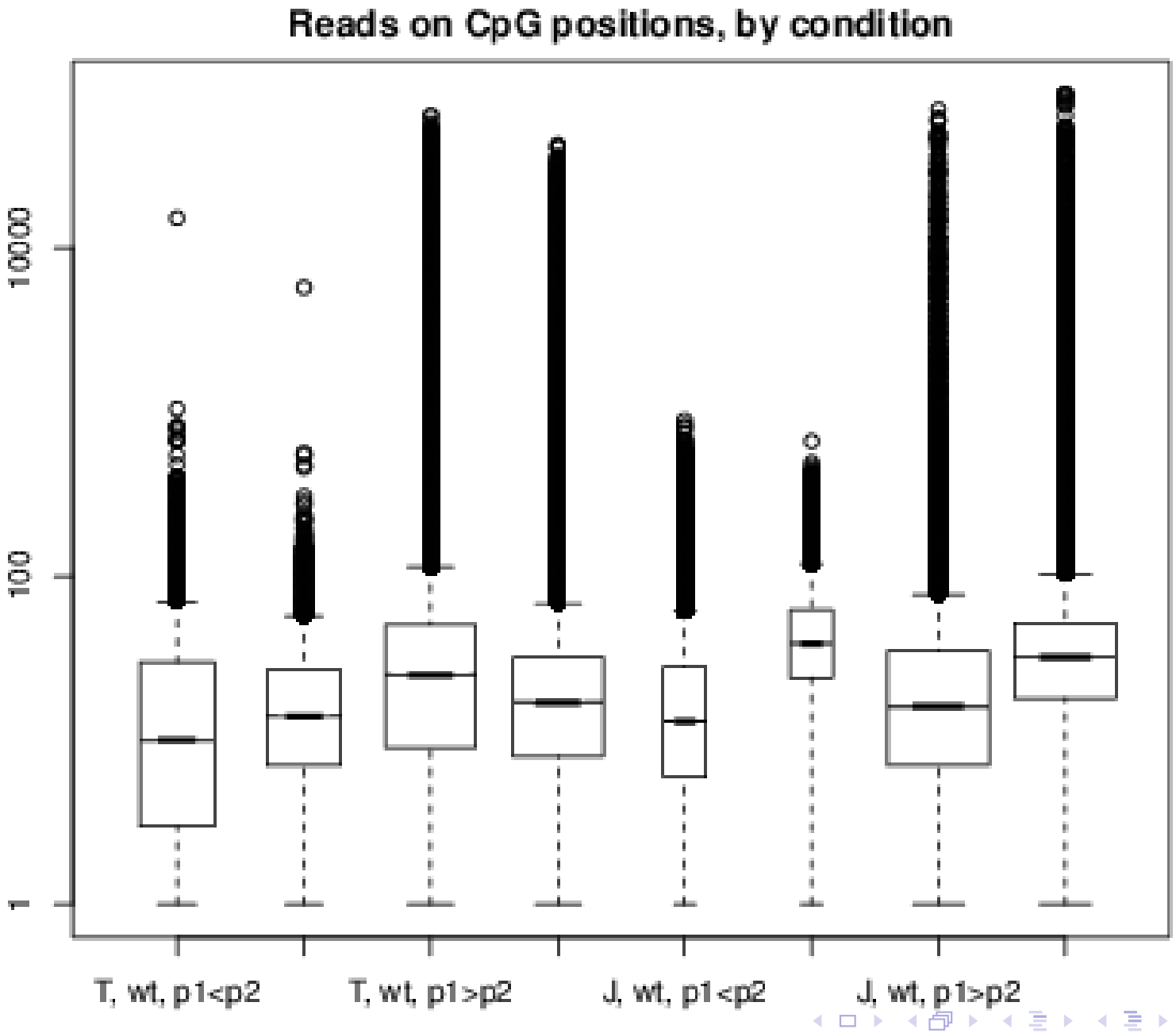
- Bismark reported a low mapping efficiency, especially for met1-3 ($\sim$44%).
- Compared to Jacobsen's data, there were many more positions with very few reads (graph on next slide).

I am currently using another trimming programme, Trim Galore!, and experimenting with other Bowtie2 parameters, hoping to improve coverage.

Problems with the data



Problems with the data



Notation

- Let $X = \{1, ..., L\}$ represent all of the positions of a chromosome or part of a chromosome.
- Let $I \subseteq X$ denote the positions of cytosines in the particular context that we are interested in.
- For each $i \in I$, let $n_i^{(c)}$ denote the number of reads on position i, and $m_i^{(c)}$ the number of reads in which the position is methylated, in condition c.
- Let $p_i^{(c)} = m_i^{(c)} / n_i^{(c)}$ be the methylation proportion at that point.

In this presentation we use $c = 1$ for the wild type plant, and $c = 2$ for the met1-3 mutant, but we could be comparing any two other conditions.

Testing for pointwise differential methylation

Although we are not interested in this, it is useful to consider how to test whether an individual point $i \in I$ is differentially methylated. We can do this using a Wald test, in which we consider the statistic

$$z_i = (p_i^{(1)} - p_i^{(2)}) \sqrt{\frac{\nu_i}{2p_i(1-p_i)}} \tag{1}$$

where

$$\nu_i = \frac{2}{\frac{1}{n_i^{(1)}} + \frac{1}{n_i^{(2)}}} \tag{2}$$

and

$$p_i = \frac{m_i^{(1)} + m_i^{(2)}}{n_i^{(1)} + n_i^{(2)}}. \tag{3}$$

Testing for pointwise differential methylation

- We assume that the number of reads with methylation is a binomial random variable.
- If the $n_i^{(c)}$ are fairly large, then z_i approximately follows a $N(0, 1)$ distribution. (This follows from the normal approximation to the binomial and t distributions.)
- If z_i is greater than a critical value, we call the point i differentially methylated.

(Concern: The normal approximation to the binomial is not so good if p is close to 0 or 1, as is often the case here. Another test, such as Fisher's exact test, may be more appropriate, but Fisher's exact test would take rather longer to run. The normal approximation is also poor if we have few reads, but we will be excluding these points anyway.)

Interpolation

- Methylation occurs only at cytosine positions, which are not evenly spaced out on the genome.
- Furthermore, we don't have the same number of reads on all cytosine positions. For many cytosine positions, we don't have any data at all.
- We are interested in differentially methylation in *regions*, rather than individual points. We want to consider trends in methylation in a region. It therefore makes sense to interpolate the methylation data at the cytosine positions to the whole of the genome.

Interpolation with moving averages

- One way to interpolate the data is by taking a moving average of the $p_i^{(c)}$ and the ν_i. This also helps to smooth out outliers and anomalies, whilst preserving trends in methylation proportion over large regions (where the meaning of 'large' is up to us).
- Then we can use the Wald test or Fisher's exact test on the moving averages in order to determine regions where nearby methylation data suggests differential methylation.

(At first, I also thought about calculating z-scores at each positions as before and interpolating the z-score, but this involves interpolating a statistic, rather than data. The current approach is more general as it allows the use of other tests as well.)

Interpolation with moving averages

$$\bar{\nu}_i = \frac{K \star (1_I \nu_i)}{K \star 1_I} \tag{4}$$

$$\bar{p}_i^{(c)} = \frac{K \star (1_I \nu_i p_i^{(c)})}{K \star (1_I \nu_i)} \tag{5}$$

$$\bar{p}_i = \frac{K \star (1_I \nu_i p_i)}{K \star (1_I \nu_i)} \tag{6}$$

where $\star$ represents convolution, and K is a triangular kernel.

Interpolation with moving averages

- I calculated moving averages by convolving against a triangular kernel K.
  - This can be done quickly (in $O(L \log L)$ time) using the Fast Fourier Transform algorithm. It takes $\sim$15 minutes per chromosome on the server.
- The moving average at a position is influenced by the methylation data of cytosine positions up to λbp away, but far positions have less influence than near positions.
  - I have been taking $\lambda = 100$ (see later).
  - If there are no cytosines within λ on either side of a position, then the moving averages are left undefined at those positions.
- The moving average is also weighted by ν_i; cytosine positions that are covered by more reads influence the moving average more strongly.
  - ν_i is the harmonic mean of the $n_i^{(c)}$.
  - It is small if either of the $n_i^{(c)}$ are small.
  - Positions influence the moving average strongly if they are well-covered in both conditions, but weakly otherwise.

Identifying DMRs

After calculating the moving averages, we can use the Wald test on the moving averages, and consider

$$z_i = (\bar{p}_i^{(1)} - \bar{p}_i^{(2)}) \sqrt{\frac{\bar{\nu}_i}{2\bar{p}_i(1-\bar{p}_i)}} \tag{7}$$

which we calculate at all i where the moving averages are defined. We then find the intervals where z_i exceeds a critical value, and take those as the list of DMRs.

Refining our list of DMRs

- Jacobsen *et al.* refine their list of DMRs by merging together any two where the gap between them was not more than 200bp. They also drop any interval where the difference in methylation proportion is less than a certain threshold (0.4 for CpG, 0.2 for CHG, 0.1 for CHH).
- Right now, I do the same, but I also eliminate any isolated DMR not longer than 50bp and any DMR with poor coverage (no more than 3 reads per cytosine position on average).
- Perhaps it would be better not to use such sharp thresholds for merging and elimination?
  - Rather than separating two close but genuinely distinct DMRs, a gap could appear if the z-score dips just below the critical value. Could we look at the value to which the z-score falls within this gap, and merge the regions only if the gap is small and the z-score in the gap is not too low?

Analysis of results

I tested the above method by comparing methylation for wild type and met1-3, and seeing whether the results were what Marco and Radu expected to see.

- My method identified a number of $-$DMRs, where the CpG methylation proportion was higher in met1-3 than in wt. This was surprising, since CpG methylation is meant not to exist in met1-3.
- However, in CpG, the total length of $-$DMRs is only around 0.1% of the total length of $+$DMRs. We identified 16,763 $+$DMRs but only 179 $-$DMRs. The $-$DMRs were much shorter than the $+$DMRs.
- When we looked at CHH methylation, we found 2,183 $+$DMRs and 5,830 $-$DMRs.

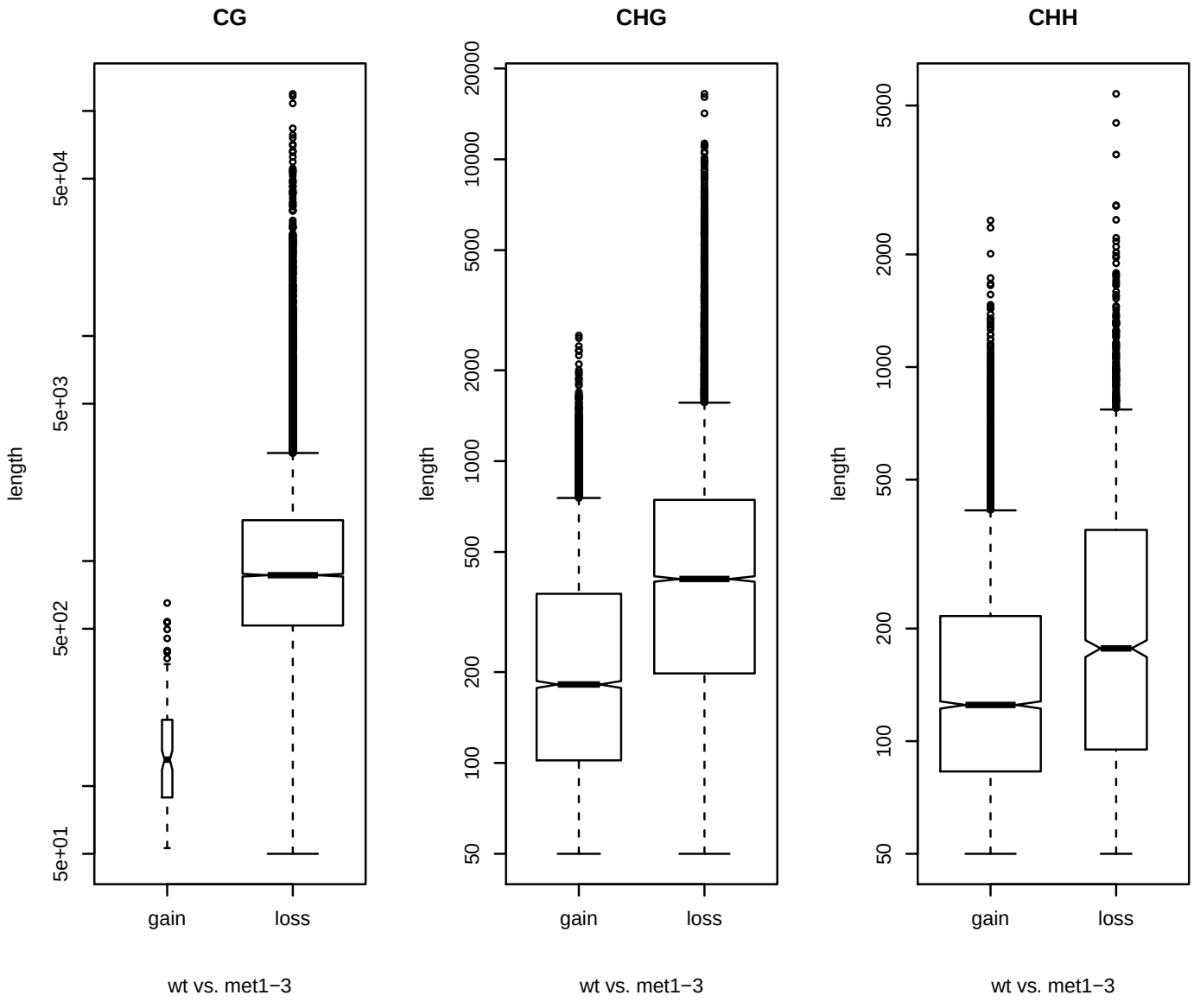
Comparing the results with Jacobsen's

For a small section of the genome, I have tried to compare the DMRs predicted by this method against the DMRs predicted by Jacobsen's, by considering:

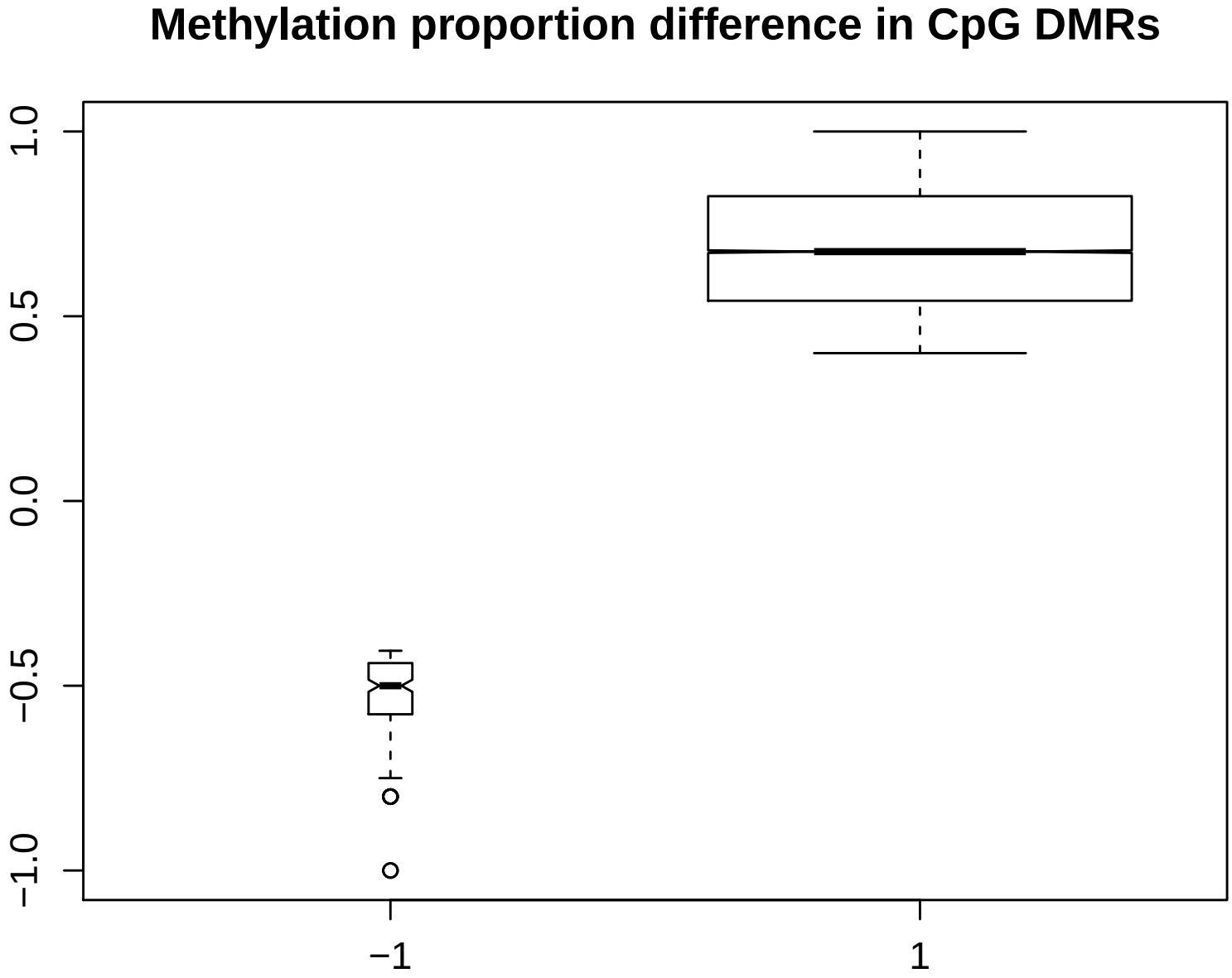
- the lengths of DMRs: running a Kolmogorov-Smirnov test on the lengths to see whether they obey a similar distribution
- the number of DMRs predicted by Jacobsen which don't overlap with any of our predictions
- the number of positions at which we disagree

but I have not done this for the whole genome yet.
We can use these measures of difference to choose what window size to take.

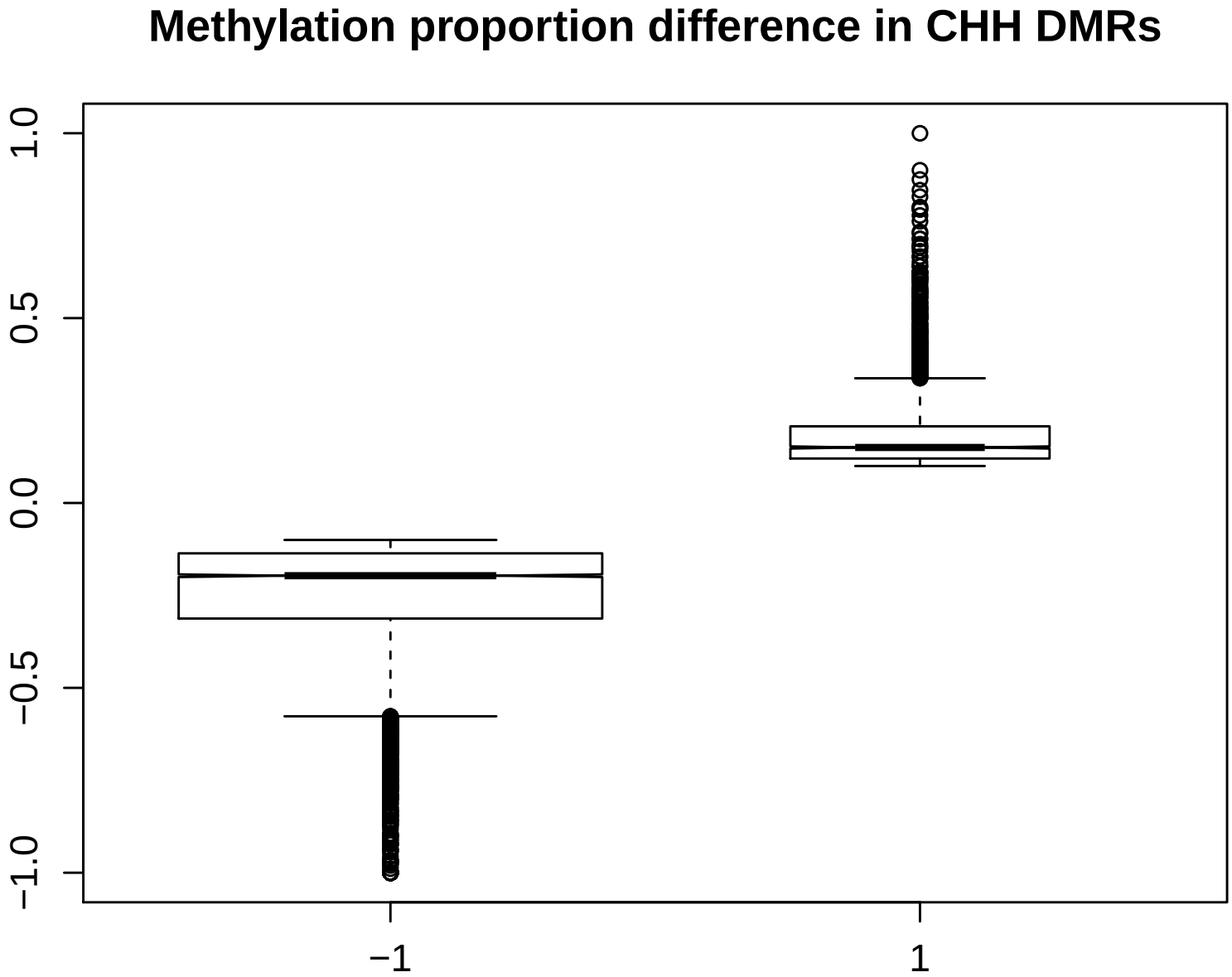
Lengths of DMRs by context



Methylation proportion difference in CpG DMRs



Methylation proportion difference in CHH DMRs



Technical details

- I wrote most of my code in R (except for the occasional perl or PHP script for preprocessing), making heavy use of the GenomicRanges package.
- DMRs are stored internally as GRanges objects with metadata (most importantly sign and context).
- With parallelisation, a genomewide analysis using this method takes between one and two hours on the SLCU server. This is very good, given that existing methods take days!

Acknowledgements

- Radu Zabet and Marco Catoni, my supervisors
- the rest of the Paszkowski Group
- Dr. Gog, for setting me up with the SLCU