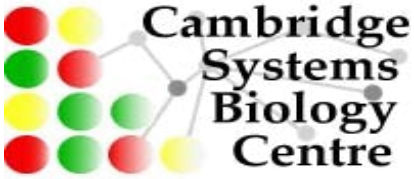

Towards a Predictive Model of DNA Cytosine Methylation in *Arabidopsis thaliana*

Jonathan Williams

Supervisors: Professor Jerzy Paszkowski,
Dr. Marco Catoni, Hajk Drost

Total Word Count: 5,944

Summary

Cytosine methylation plays a key role in epigenetic regulation of the plant genome. Unlike in animals, it is possible for the methylation status of the plant genome to be passed between generations. In particular, *Arabidopsis thaliana* mutants with loss of function for methylation are able to transmit their demethylated chromosomes. Interestingly, once crossed with wild-type plants, methylation is not always immediately regained, often taking many generations, if returning at all. The time frame over which this happens varies across the genome. Presented here is an attempt to predict the ordering in which regions regain methylation. This was approximated by a partial loss of methylation function mutant, where regions of the genome show differential levels of relative methylation compared to wild-type individuals. It has been observed that highly repetitive elements of the genome tend to be more heavily methylated, with methylation in all contexts of cytosines, whereas genes tend to be methylated primarily via cytosines found in the symmetric CG context and take longer to regain methylation. As such, a collection of models are presented, considering local tiles of the *A. thaliana* genome, scoring them based upon their CG context cytosine content and how repetitive they are in the genome, to predict their expected order of remethylation. The models achieve accuracy better than random guessing, confirming that these measures appear to influence the rates of remethylation. However, the models fail to predict the majority of regions accurately and so further work is needed to refine this model.

Table of Contents

Abbreviations

TEL – Transposon-like methylation

GEL – Gene body-like methylation

MET1 – DNA methyltransferase 1

Introduction

DNA cytosine methylation in *Arabidopsis thaliana*

Cytosine methylation is an important and well-studied epigenetic mark in eukaryotic organisms, which can have consequences for the behaviour and regulation of elements of the genome (Jones, 2012). In *Arabidopsis thaliana* the presence of cytosine methylation has been linked to differential rates of transcription (Zilberman et al, 2007) and is also thought to play a role in silencing of transposable elements in the genome (Rabinowicz et al., 2003; Tompa et al., 2002; Zhang et al., 2006). Such methylation is present throughout the genome and occurs in three contexts: CG, CHG and CHH (where H is A, C or T) (Cokus et al., 2008; Stroud et al., 2014). Hereafter, the CG context will be referred to as CpG for clarity.

De novo methylation in plants is achieved through the use of small RNAs which target cytosine methylation in all contexts for homologous sequences of DNA (Wassenegger, Heimes, Riedel, & Sanger, 1994). This RNA-directed DNA methylation (RdDM) pathway uses the components of the RNA interference pathway in combination with two plant specific DNA polymerases, POL IV and POL V (M. A. Matzke & Mosher, 2014; M. Matzke et al., 2009; Pickford & Cogoni, 2003; Zilberman et al., 2004). In addition to this, each context of cytosine methylation is maintained by their own separate pathway (Law & Jacobsen, 2010). Cytosine methylation in a CpG context is the most common context for methylation in plants and is maintained by DNA methyltransferase 1 (MET1) (Goll & Bestor, 2005; Lister et al., 2008). CHG context methylation is maintained primarily through chromomethylase 3 (CMT3) and CHH context methylation is maintained through the RdDM pathway, as well as by CMT3 at

a low level (Bartee, Malagnac, & Bender, 2001; Chan, Henderson, & Jacobsen, 2005; Johnson et al., 2007).

The role of DNA cytosine methylation in the *A. thaliana* epigenome

The need for maintenance of cytosine methylation arises from the very dynamic and adaptable nature that this epigenetic mark possesses. Interference in these pathways can give rise to phenotypic differences, referred to as epimutations, where only the epigenome of a plant has been altered (Kakutani et al., 1996). This can also give benefit to plants, where previously epigenetically silenced transposable elements can be activated under stress by alteration of their methylation status enabling potentially rapid alteration of the host genome. (Mirouze & Paszkowski, 2011; Miura et al., 2001; Tsukahara et al., 2009). Unlike in mammals, where nuclear reprogramming results in the removal of most cytosine methylation between generations, plants are able to transfer the majority of cytosine methylation across generations (Heard & Martienssen, 2014). In plants, the majority of CpG and CHG context methylation is retained in the germline. By contrast CHH methylation is mostly lost, however the presence of 24nt siRNA in the seed is able to return methylation via targeting to many elements of the genome that lose this CHH methylation (Calarco et al., 2012). As a result, transgenerational inheritance of epigenetic markers is possible in plants. This gives rise to the concept of an epiallele; an allele that is identical in terms of genetic sequence, but which has an altered but stable state of cytosine methylation which can give rise to altered phenotypes (Paszkowski & Grossniklaus, 2011; Weigel & Colot, 2012).

Further transgenerational epigenetic behaviour has been studied by breeding plants with a mutation that interferes with the methylation maintenance pathways. Such mutants display hypomethylation or complete loss of cytosine methylation. Once crossed with individuals with wild-type methylation machinery, the methylation levels do not return immediately to wild-type levels and in many cases the methylation levels increase over the course of many generations of crosses, with different regions of the genome returning at different rates (Kakutani et al., 1999; Saze, Mittelsten Scheid, & Paszkowski, 2003).

The mechanism through which the gradual return of methylation occurs in partial and full loss of methylation pathway mutants is not currently known. The work of Dr. Catoni proposes that the mismatch in rates of return of methylation could be related to the underlying genomic features of the regions in question. It has already been observed that highly repeated regions of plant genomes often tend to be more heavily methylated, and given the small RNA mediated pathway for de novo methylation, it is likely that more repetitive elements of the plant genome would be more targeted for methylation, especially after global methylation reduction (Law & Jacobsen, 2010; Selker, 1999; Teixeira et al., 2009). In addition, the maintenance of CpG context methylation between generations has been suggested to be key to stability of epigenetic inheritance (Mathieu et al., 2007; Saze et al., 2003). Hence, a mutation that affects the CpG context methylation would be more likely to cause long-term loss of methylation in regions with relatively high CpG methylation in wild-type individuals.

The role of the CpG context and repetition in cytosine methylation

It has been established that in the *A. thaliana* genome, regions coding for genes tend to have high methylation in CpG contexts, but low methylation in other contexts (Takuno & Gaut, 2013; Tran et al., 2005). By contrast, transposable elements appear to be more heavily targeted across all contexts of cytosine methylation (Tompa et al., 2002). These differences between elements of the genome have led to definitions used by Dr. Catoni and the Paszkowski group. Regions of a plant genome that have methylation in the CpG context, but not in the CHG or CHH context are referred to as gene body-like (GEL) elements. Elements that have methylation in all three contexts are similarly referred to as transposon-like (TEL) elements. As a basic definition, a cut off of methylation of at least 5% of cytosines in a given context for a region was used. In working with these definitions and with mutants affecting the MET1 pathway, it was noted that in general GEL regions tend to be regions where methylation takes a long time to return, if at all, in crosses between mutants and strains with a working MET1 pathway. By contrast, TEL regions tend to have a wider spread of rates of methylation regain. Some regions behave almost like GELs and are slow to regain methylation, and are referred to as ETELs (for epiallelic TELs as they can create stable unmethylated epialleles), whilst other regions have almost immediate recovery of methylation and are referred to as RTELs (for reversible TELs as they can return to their original methylation state). As a proxy to define these regions, the relative level of methylation under the partial loss of function *met*1-1 mutation is used. TELs retaining at least 80% methylation are labelled as RTELs, and those retaining less than 5% are ETELs. Figure 1 shows a decision tree to help understand this classification.

In studying the behaviour of ETELs and RTELS it was noted that ETELS tend to contain more cytosines in a CpG context (regardless of methylation) and that RTELS tend to have sequences that are very repetitive in the genome. It was proposed that repetition and CpG methylation could be key in the mechanism leading to differential rates of methylation return. To test this, it was decided to attempt to build a model to predict the rate of methylation return in differing regions of the *A. thaliana* genome based upon these features. The potential for CpG methylation was modelled simply by the count of CpG context cytosines, as this measure would require no assumptions about the wild-type methylation status. As a first approach to measuring repetition, BLAST was considered a valid measure. Here presented is an attempt to build such a model. What results has limited predictive power, but clearly shows that these two measures factor in to the relative rates of methylation return.

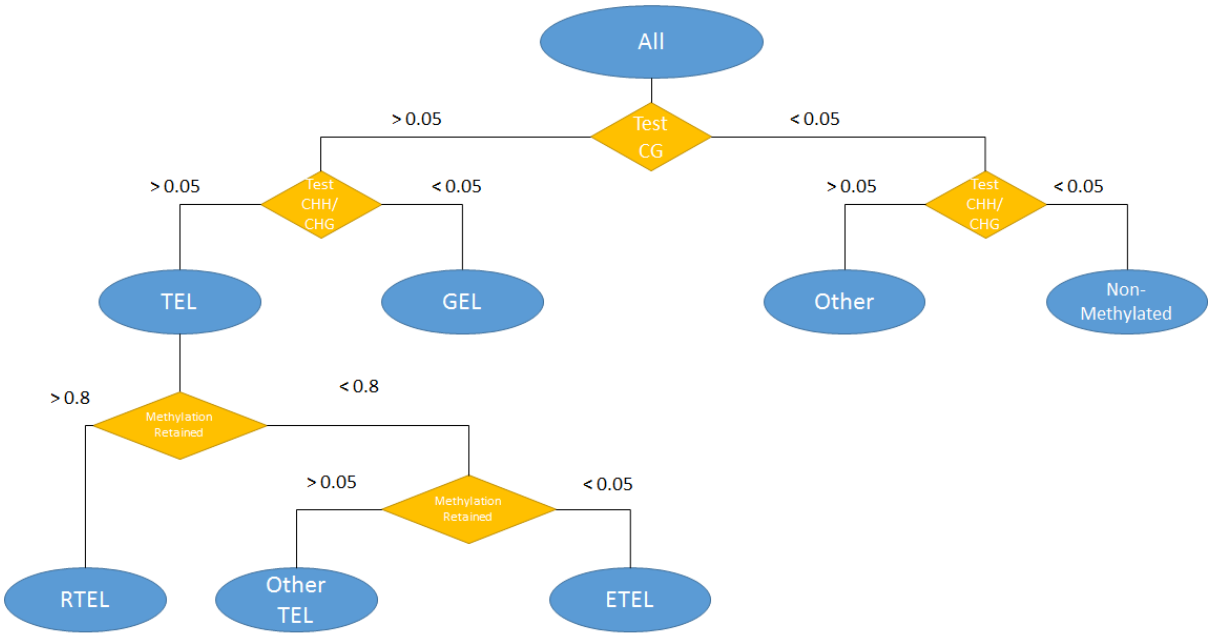

Figure 1: Decision tree for classifying regions of a genome as GELs and TELs. For CG, CHG and CHH methylation, the score is the proportion of cytosines in that context which are methylated. The methylation retention is scored as the relative methylation under the *met*1-1 mutation when compared to wild-type.

Results

Raw Data

In order to develop a model that could predict the relative order of return of DNA cytosine methylation in different regions of the *A. thaliana* genome required a data set with DNA cytosine methylation measured under different mutations. The data set that was used to build the model was taken from the work of Dr. Marco Catoni. Here the genome was summarised in overlapping, sliding tiles. These tiles started at the first base pair of each chromosome and were 200 base pairs long. Each subsequent tile was started at an offset of 50 base pairs, resulting in the vast majority of the genome being covered in 4 different tiles (with the ends of chromosomes being the only exceptions). Within each tile, the number of cytosines in each of the contexts, CpG, CHG and CHH were reported.

In order to measure the methylation of the cytosines in each tile, the genomes of various mutants had been sequenced using bisulphite sequencing. In particular, multiple MET1 mutants had been sequenced, including *met*1-1 partial loss of function and *met*1-3 complete loss of function mutants (Kankel et al., 2003; Saze et al., 2003). Such sequencing enables single nucleotide resolution measurements of cytosine methylation (see Methods and Materials). The results of this sequencing were summarised in totals of both methylated and unmethylated cytosine reads for each 200bp tile, enabling the calculation of the proportion of methylated cytosines. This proportion of methylation was used as the measure of methylation for each tile.

In addition to the cytosine methylation data, each tile had been compared to the rest of the *A. thaliana* genome using BLAST. A score was reported for each tile, representing the number of hits that it returned via BLAST considered against the rest of the genome. This BLAST score was used as a way of measuring how repetitive each tile was within the whole genome.

Measuring Relative Levels of Methylation

In order to be able to predict the order in which regions of the *A. thaliana* genome would become methylated again once crossed with a wild-type individual, if at all, after alterations to the MET1 pathway, it was necessary to calculate a score that reflected this process from the available bisulphite sequencing data. For simplicity it was decided to use the relative proportions of cytosine methylation in CpG contexts under the *met*1-1 mutation, when compared to wild-type levels for each tile. This is a mutation that causes hypomethylation, especially in the CpG context, but not complete loss of methylation. It has been noted in literature (Kankel et al., 2003), as well as witnessed in the work of Dr. Catoni, that the relative hypomethylation under the *met*1-1 mutant can be transferred across generations and that in crosses with wild-type the methylation can return, and hence this mutation has behaviour similar to that of epialleles under complete loss of MET1 mutations when later crossed with wild-type plants (Saze et al., 2003). Figure 2 shows the distribution of the relative CpG context cytosine methylation for regions of the *A. thaliana* genome that meet the previously set out criteria to be GELs and TELs respectively, confirming that this measure recapitulates their expected behaviours to a reasonable extent.

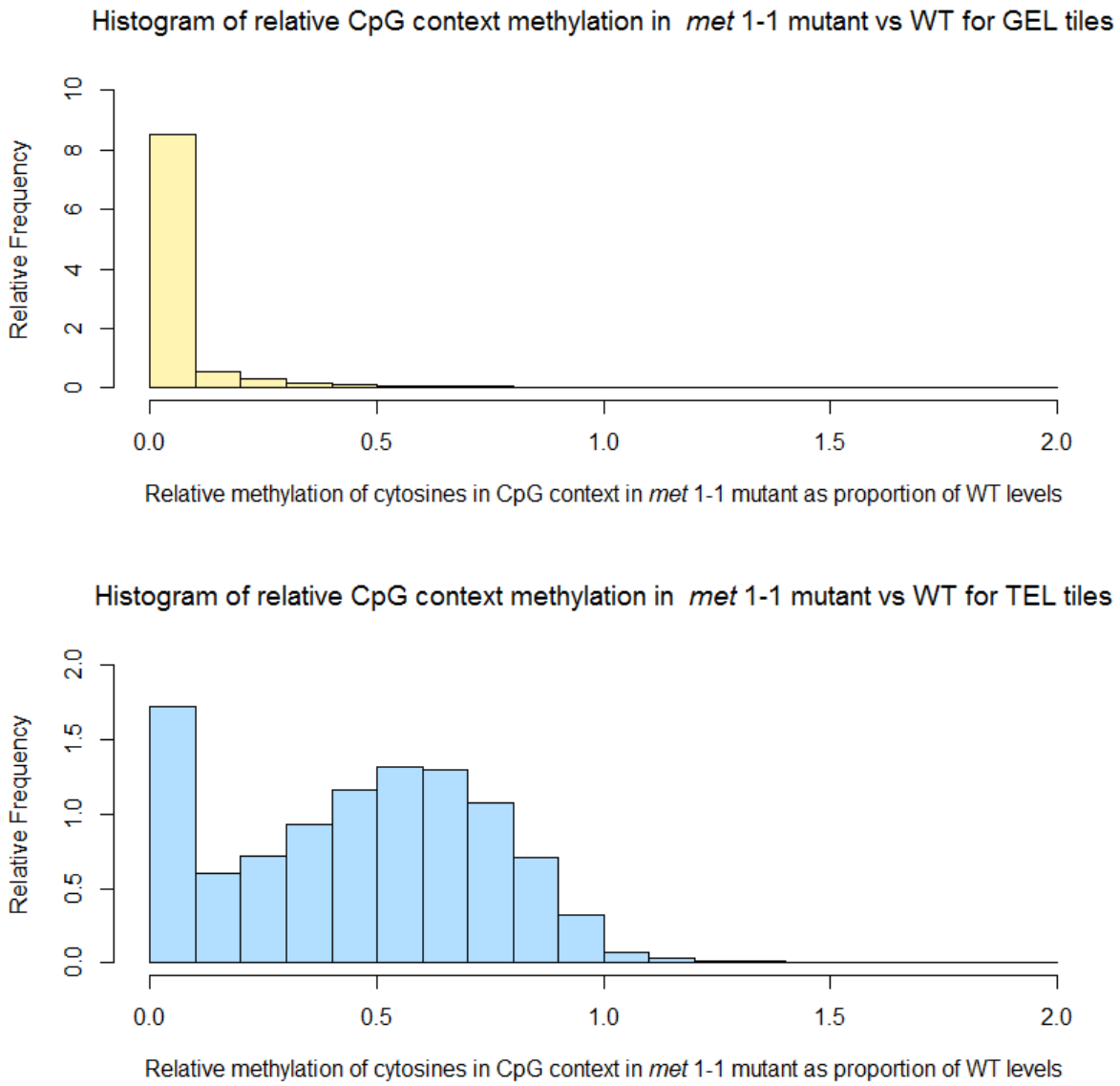

FIGURE 2: Histograms of relative methylation for tiles of the *A. thaliana* genome meeting criteria of GEL and TEL definitions respectively.

Displayed are the distributions of the relative proportions of cytosines in CpG context that are methylated for each 200bp tile under *met*1-1 as a fraction of the proportion methylated in WT. These are shown for tiles that meet the criteria for GELs and TELs respectively. It is clear that the tiles labelled as TELs have generally higher scores under this measure, showing that this measure can be reasonably used as a score for the relative propensity of each tile to regain or retain methylation. (Relative methylation scores of over 2 have been removed from the figure for clarity, though they make up a very small proportion of the total data set)

Regression Models

A selection of regression models were fitted to the data to explore the power of the CpG context cytosine count and BLAST score in predicting the relative methylation in *met*1-1 mutants. Table 1 displays a summary of the most pertinent models and their results in predicting the relative methylation score. It is clear from this table that direct prediction through some form of general linear model was unlikely to yield success. Even the model combining CpG context cytosine count and BLAST score for each tile (including interaction) produces an R-squared value of only 0.0131. Hence this model can only explain a very small fraction of the variation in relative methylation.

Model	Intercept	CpG Count Coefficient	p-value	BLAST score Coefficient	p-value	Interaction Coefficient	p-value	R^2
CpG Count only	0.328	0.0238	<2e-16	N/A	N/A	N/A	N/A	0.0124
BLAST score only	0.602	N/A	N/A	-2.21e-4	<2e-16	N/A	N/A	8.55e-5
CpG count & BLAST score	0.317	0.0250	<2e-16	9.05e-4	<2e-16	-1.28e-4	<2e-16	0.0131

Table 1: Summary of linear regression models for predicting *met*1-1 relative methylation for each tile. Models are included for regression based on CpG context cytosine count, BLAST score and a combination of both for each tile. The models all score very poorly when R-squared is measured. Where values were missing or the relative methylation was not well-defined, tiles are omitted.

The most striking result from the regression models is the incredibly small effect including BLAST score has on improving the accuracy of the model compared to one based solely on the count of cytosines in a CpG context. This suggested that using BLAST hits for each tile was not an effective way of quantifying how repetitive the sequence within that tile is when compared to the whole genome. BLAST score had appeared to be a promising way to distinguish between the classes of elements, such as GEL and TEL, that the model hoped to be able to identify. This is shown by the apparent zoning seen in figure 3. However, this misleading as the scatter plot doesn't

represent the true distributions of the BLAST score. When plotted as a smoothed density plot, the issue with BLAST is made clear (Figure 4). It is clear that BLAST scores are mostly clustered at very low values and hence a better measure of how repetitive each tile was needed in order to better differentiate the tiles.

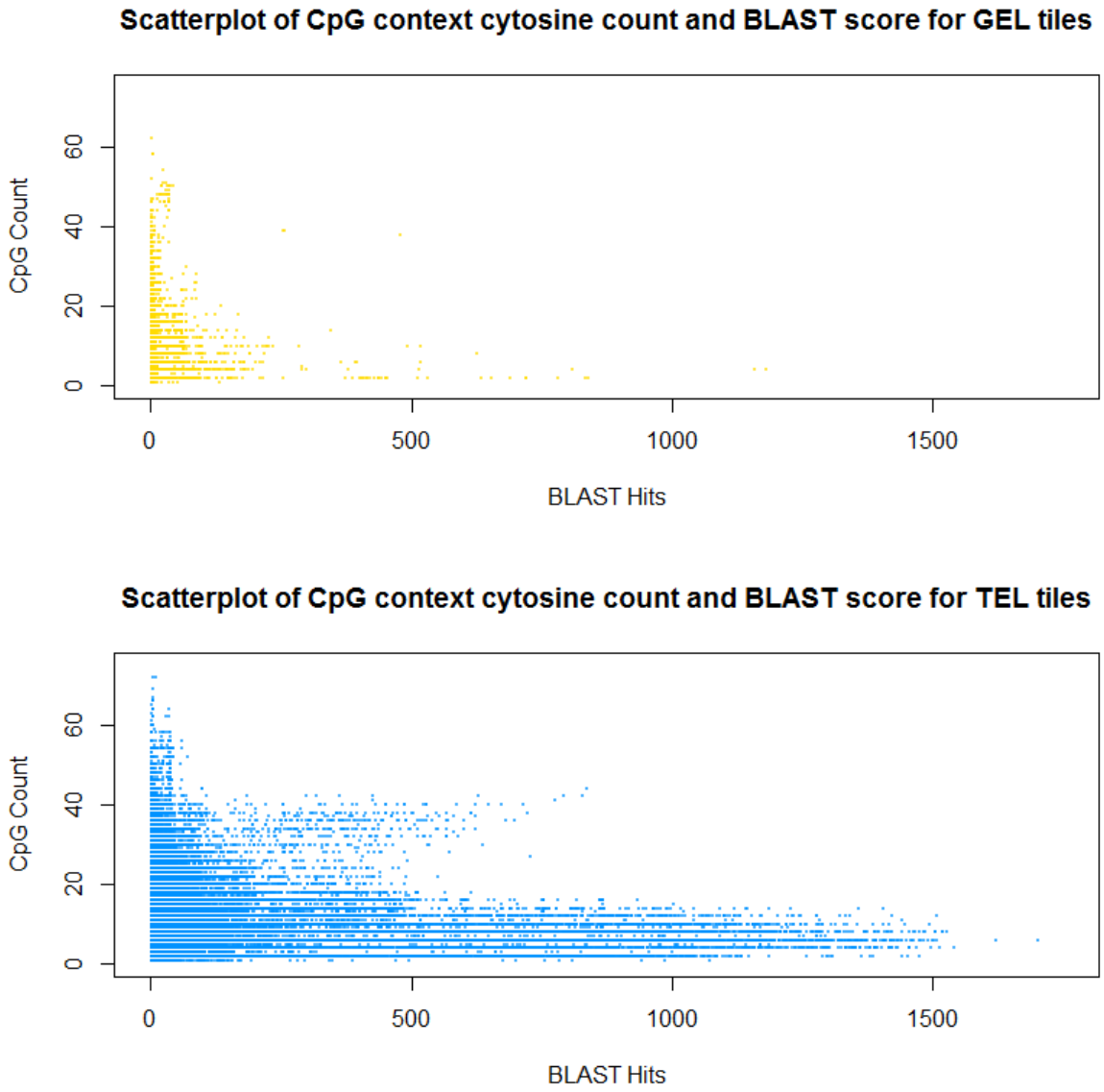

Figure 3: Scatterplots of CpG context cytosine count and BLAST score for GEL and TEL tiles. There appears to be a trend for a greater number of high BLAST scores for TEL tiles as opposed to GEL tiles. This is consistent with previous studies linking repetitive sequences with greater occurrence of methylation, as expected for TELs

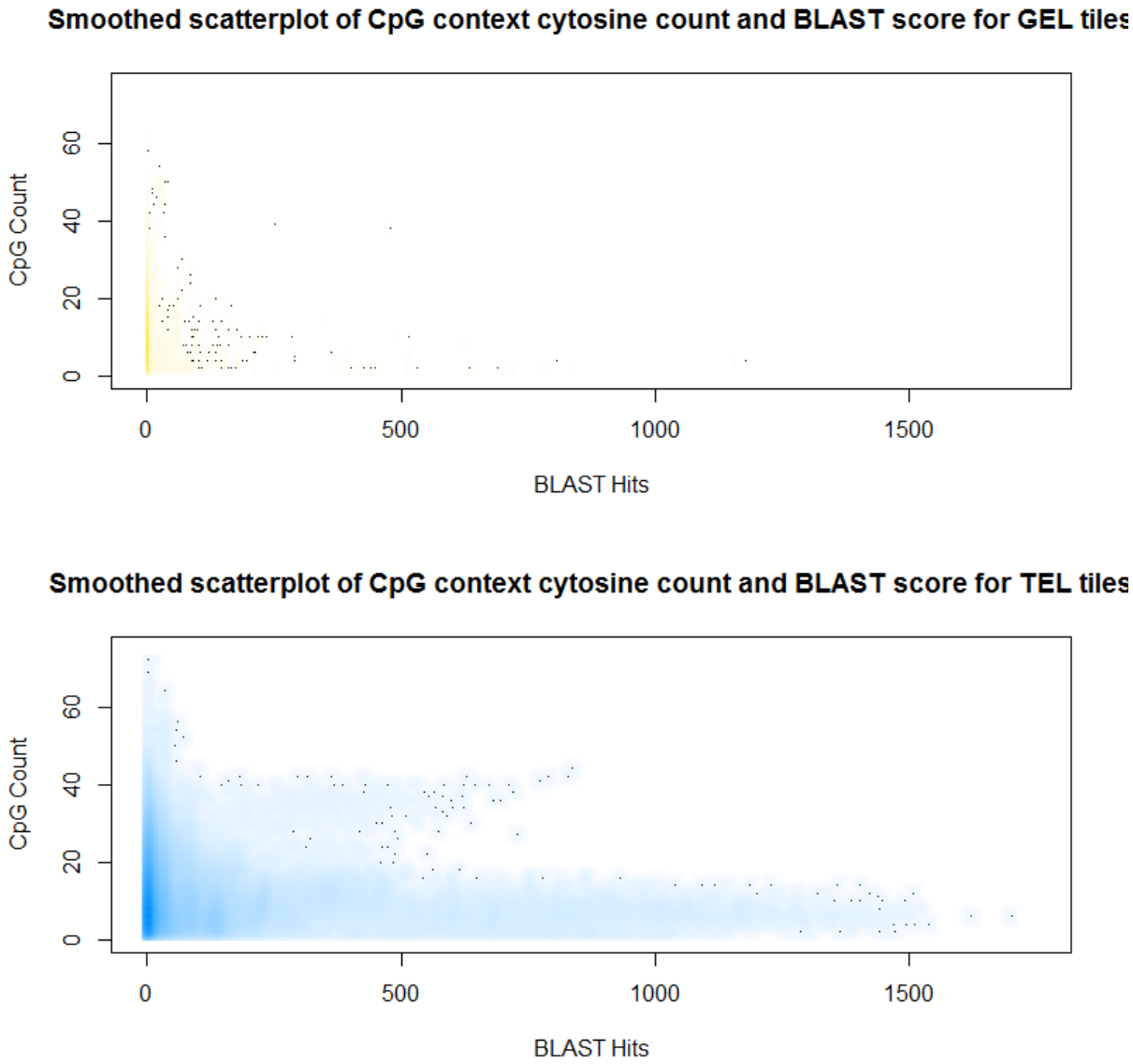

Figure 4: Smoothed scatterplots of CpG context cytosine count and BLAST score for GEL and TEL tiles. With smoothing added to the scatterplots it becomes clear that both GEL and Tel elements have a dense clustering of BLAST score in the low values.

Mappability as a measure of repetitiveness

To deal with the difficulties of differentiating the repetitiveness of tiles by using BLAST, a mappability score was calculated for each tile. Mappability scores how well a given sequence is able to be uniquely mapped to a given genome, with a score of 1 meaning the sequence is unique and lower scores indicating some degree of repetition in the genome in question (see Methods and Materials). Hence a mappability score can be used to quantify the repetitiveness of a given sequence, interpreting low scores as

representing high repetitiveness, and scores close to one as representing near unique sequences.

The mappability score for each tile was calculated by first computing a mappability score for each nucleotide within the tile. This was scored by selecting the sequence centred on the given nucleotide for a set number of base pairs. This was then scored against the *A. thaliana* genome with a specified number of mismatches allowed. The average of these single nucleotide scores was then calculated for each 200bp tile to give a score for the whole tile. This approach avoided the issue that BLAST scoring faced whereby the 200bp tile sequence was being scored as a whole. The averaging approach enables sub-sequences within each tile to be tested and contribute to overall repetitiveness score for each tile.

Initially it was decided to try running the routine with a width of 200bp for the window to be compared for each nucleotide, as this was comparable to the BLAST methodology. In this case it was chosen to allow 8 mismatches. In addition to this the same pipeline was used to calculate scores where the window for each nucleotide was set to be 20bp wide, this was selected as it is similar to the length of many small RNAs (Heard & Martienssen, 2014). Here the same score was calculated for allowed mismatches of 0,1,2 and 3. In order to decide which score to use moving forward, a linear regression model was once again fitted as before and the scores a presented in table 2. Based on the values for R-squared it was decided to use the mappability score given by using a window of 20bp width with 1 mismatch allowed. This choice was made as it gave the same value for R-squared as allowing more mismatches, but it

was more restrictive should there be a sequence recognition mechanism at play than allowing more mismatches. Hence it was more likely to reflect the affinity for any such mechanism to the sequences being studied.

As can be observed in figure 5, the mappability score chosen shows much a much clearer spread for the data when compared to the smoothed scatter plots in figure 4. This enables a better differentiation between tiles based upon their repetitiveness. From figure 5, it is also observed that the distributions for GEL and TEL tiles appear qualitatively different, with a greater proportion of TEL tiles having a high count of cytosines in CpG contexts. Additionally, a greater proportion of TEL tiles score close to zero for mappability, indicating a higher proportion of these tiles being highly repeated in the *A. thaliana* genome.

Mappability Routine	Intercept	CpG Count Coefficient	p-value	Mappability Coefficient	p-value	Interaction Coefficient	p-value	R^2
200bp 8 mismatch	0.719	-0.0251	<2e-16	-0.425	<2e-16	0.0525	<2e-16	0.0153
20bp 0 mismatch	0.828	-0.0357	<2e-16	-0.576	<2e-16	0.0676	<2e-16	0.0179
20bp 1 mismatch	0.720	-0.0213	<2e-16	-0.519	<2e-16	0.0580	<2e-16	0.0192
20bp 2 mismatch	0.720	-0.0212	<2e-16	-0.519	<2e-16	0.0580	<2e-16	0.0192
20bp 3 mismatch	0.720	-0.0212	<2e-16	-0.519	<2e-16	0.0580	<2e-16	0.0192

Table 2: Summary of linear regression models for predicting *met*1-1 relative methylation for each tile using mappability. The R-Squared values are still low, but compared to those in table 1 are an improvement. Beyond 1 mismatch it appears that allowing more freedom has little effect on results for windows of 20bp. Where values were missing or the relative methylation was not well-defined, tiles are omitted.

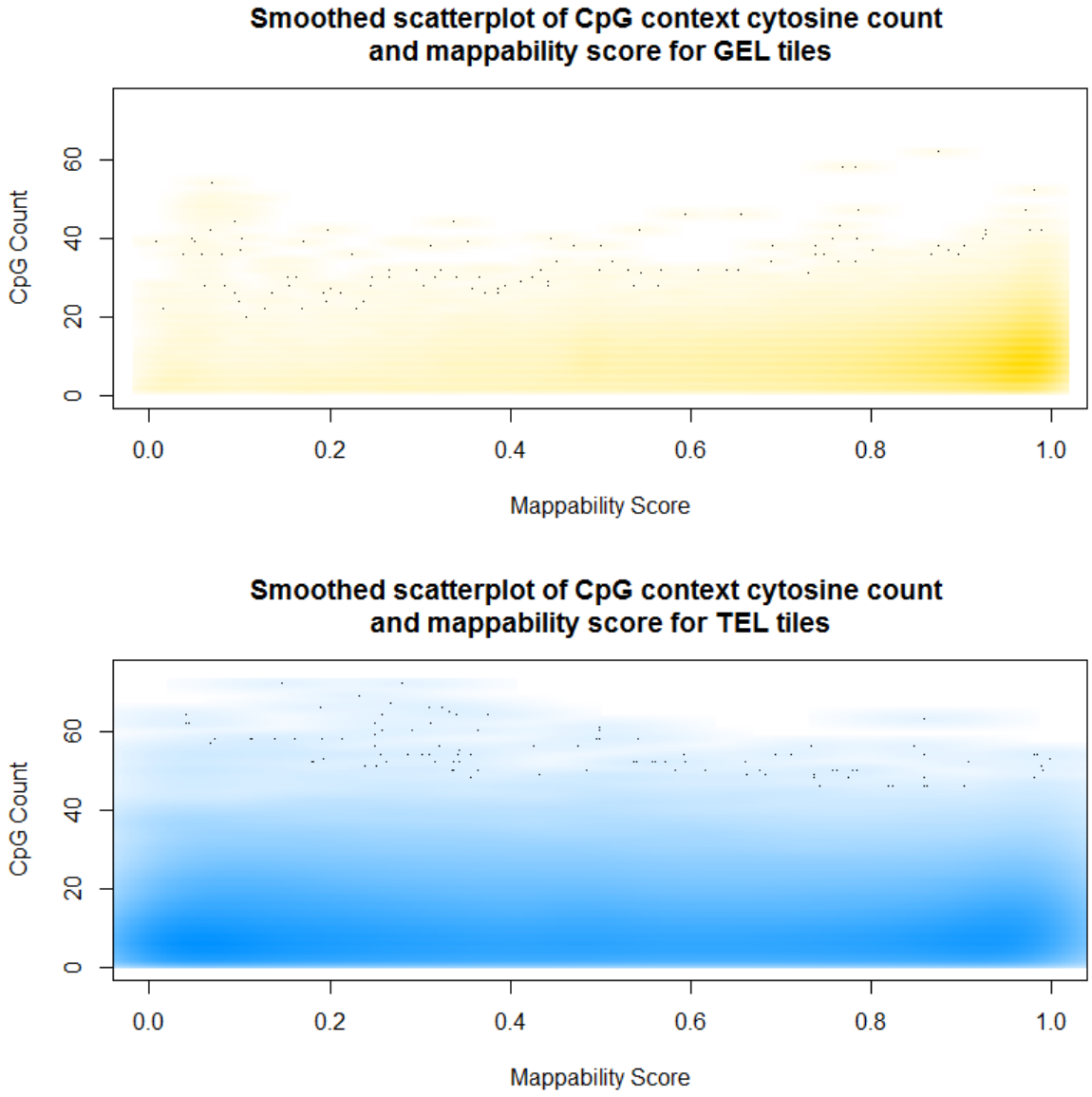

Figure 5: Smoothed scatterplots of CpG context cytosine count and mappability score for GEL and TEL tiles. When BLAST hits is replaced by the mappability routine as described (20bp window, allowing 1 mismatch) the distributions of both GEL and TEL tiles becomes better spread, enabling better differentiation of repetitiveness of each tile.

Moving from quantitative prediction to classification

The results in table 2 show that any general linear model is likely to struggle to produce meaningful predictions for the relative methylation of cytosines in *met*1-1 mutants when compared to wild-type individuals. This is clear from the very low values for R-squared in each simple linear regression model showing that there is too much noise in the data for the CpG context cytosine count and mappability alone to be able to give quantitative predictions. Though not included here, various polynomial models were fitted with linear regression and there was no significant improvement on the values of R-squared. As a result, a different approach was sought in order to be able to produce meaningful predictions.

At this point the data used as the basis for the model was filtered so as to remove missing values. In particular, for tiles where there were no cytosines in a CpG context it was not possible to quantify relative methylation under the *met*1-1 mutation as used in this model. Additionally, tiles which show no methylation in wild type individuals were removed as relative methylation is not well-defined for such tiles. This filter was applied so as to focus on classifying tiles with valid quantifications for the variables of interest.

The motivation behind building a predictive model for DNA cytosine methylation was to be able to predict how regions of the *A. thaliana* genome would behave, in terms of this methylation, following mutations that interfere with the pathways involved in setting up and maintaining the methylation. Given that a direct quantitative prediction of the exact changes under *met*1-1 mutation through any linear model method was unlikely to succeed for the model as presented, it was decided to focus on developing a

qualitative model. This would be in line with the motivation of this investigation; being able to group the tiles into sets that show similar behaviour under the *met*1-1 mutation based upon the CpG context cytosine content and repetitiveness of each tile would still serve to validate the principles of the model.

To see if this was possible, various clustering schemes were applied to the data from the tiles, in particular the k-means and multiple forms of the fuzzy c-means algorithms were applied (see materials and methods). These algorithms were given the CpG context cytosine content and all 5 mappability scores (for all of the combinations of window length and mismatches as per table 2). In all cases, whilst clusters were highlighted, where many tiles shared similar scores across these variables, the distributions of relative methylation under the *met*1-1 mutations in each cluster did not differ in a qualitatively observable way. As such it wasn't possible to highlight distinct regions on plots such as those shown in figure 5 which were notable for being areas of particularly high, nor low, relative methylation.

Classification by quantile of relative methylation

With the difficulties in achieving a classification model under an unsupervised learning approach, a model based on supervised learning was instead developed. To maintain better generality for the model it was decided to create classes based on the quantiles for the relative methylation score under *met*1-1 mutation. Whilst a model could be built where tiles were classed based upon whether they were GELs or TELs, these classes would depend heavily on the definitions of GELs and TELs and hence the differences

highlighted would likely struggle to generalise to predicting if a new tile fell under one of these classes.

As can be seen in figure 6, when the tiles are stratified by the deciles of relative methylation there appear trends in the distributions of the CpG context cytosine counts and mappability scores for these tiles. In particular, the mid-range of the deciles appear to have a particular enrichment for tiles with low mappability score as well as higher deciles tending to have a greater proportion of tiles with higher counts of cytosines in a CpG context.

In order to quantify if the distributions for the CpG context cytosine count and mappability differed significantly between the quantiles of relative methylation, the non-parametric Kolmogorov-Smirnov test was applied (see Materials and Methods). In order to enable this test to be applied, a total of 5 quantiles were used. This resulted from the large number of tiles which have complete loss of methylation. These make up roughly 20% of the total data, once filtered. Hence, in calculating quantiles, when more than 5 are desired, these tiles must be ignored and added to the bottom quantile, or else the boundary for the lowest quantile remains zero. This skews the sizes of the quantiles.

The results of the Kolmogorov-Smirnov tests are summarised in tables 3 & 4, and the distributions of the variable for each quantile are represented by their empirical cumulative distribution functions in figure 7 (see Materials and Methods). Here it is clear that stratification by relative methylation produces statistically significantly different distributions in the variables being used in the model. This results verifies

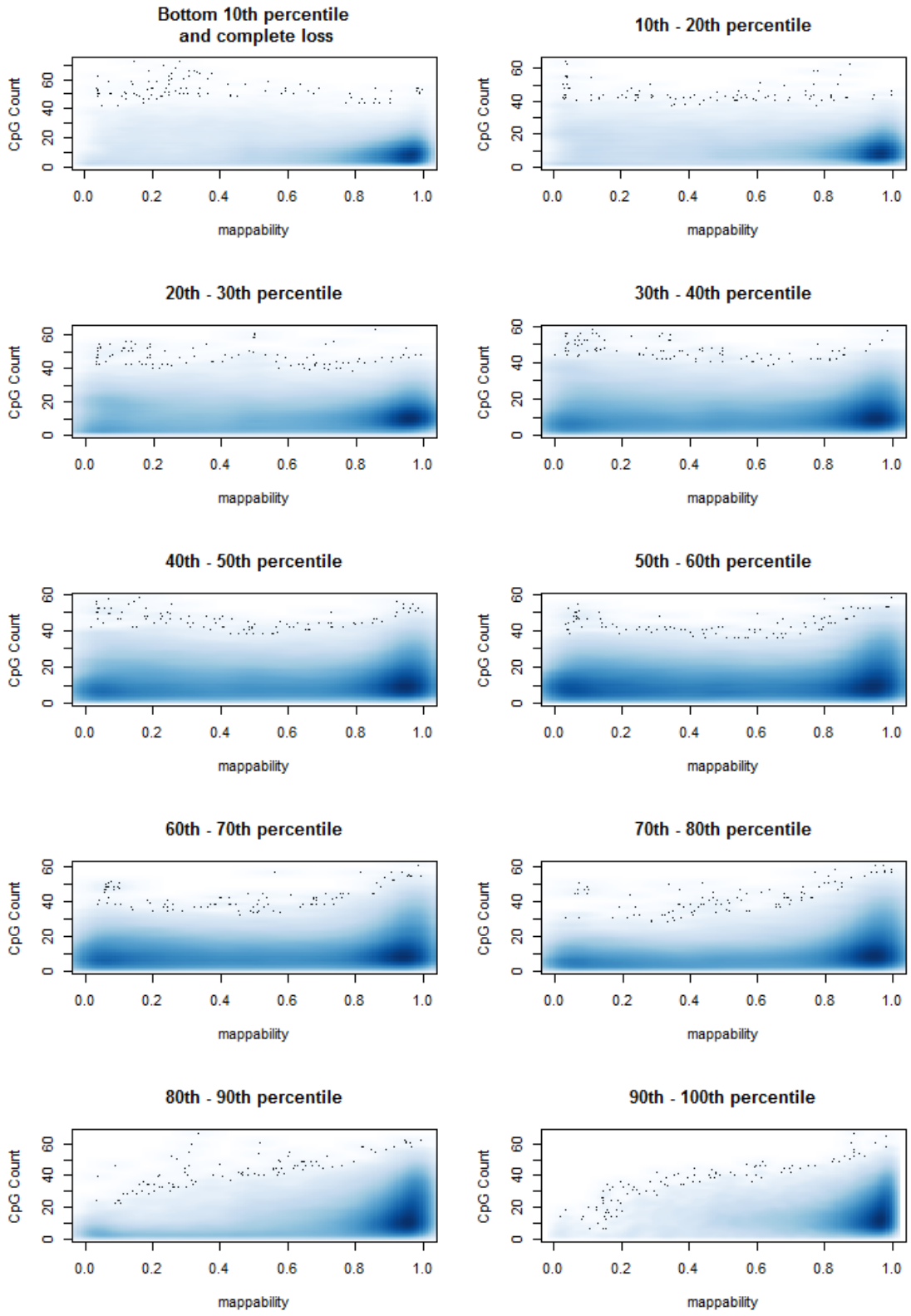

Figure 6: Smoothed scatterplots of CpG context cytosine count and mappability for tiles stratified by percentiles of relative methylation under *met*1-1 mutation. Here it is clear that mappability in particular varies a lot across regions of differing relative methylation. It can also be seen that regions with low repetition (high mappability) tend to have a wider spread of CpG count for regions of higher relative methylation. For computational simplicity, regions of complete loss of methylation were ignored in calculating quantiles and are all included in the first quantile (top left)

the principles underlying the idea upon which the model is based. Namely, that CpG context cytosine count and repetitiveness of regions of the *A. thaliana* genome do differ between regions of differing relative methylation.

Quantile	Comparison Quantile	Kolmogorov-Smirnoff Test Value	p-value
0%-20%	20%-40%	0.126	<2e-16
0%-20%	40%-60%	0.131	<2e-16
0%-20%	60%-80%	0.073	<2e-16
0%-20%	80%-100%	0.316	<2e-16
20%-40%	40%-60%	0.014	<2e-16
20%-40%	60%-80%	0.074	<2e-16
20%-40%	80%-100%	0.194	<2e-16
40%-60%	60%-80%	0.071	<2e-16
40%-60%	80%-100%	0.184	<2e-16
60%-80%	80%-100%	0.245	<2e-16

Table 3: Results of Kolmogorov-Smirnov tests for CpG Cytosine Count when stratified by 5 quantiles for relative methylation. The test statistic measures the largest difference between the empirical cumulative distribution functions. Here it is clear that the distributions differ between quantiles, however, in many cases this score is low.

Quantile	Comparison Quantile	Kolmogorov-Smirnoff Test Value	p-value
0%-20%	20%-40%	0.279	<2e-16
0%-20%	40%-60%	0.400	<2e-16
0%-20%	60%-80%	0.310	<2e-16
0%-20%	80%-100%	0.023	<2e-16
20%-40%	40%-60%	0.124	<2e-16
20%-40%	60%-80%	0.048	<2e-16
20%-40%	80%-100%	0.301	<2e-16
40%-60%	60%-80%	0.090	<2e-16
40%-60%	80%-100%	0.421	<2e-16
60%-80%	80%-100%	0.332	<2e-16

Table 4: Results of Kolmogorov-Smirnov tests for mappability when stratified by 5 quantiles for relative methylation. Again it is clear that the distributions differ between quantiles. The statistic scores are generally higher than in the CpG count tests.

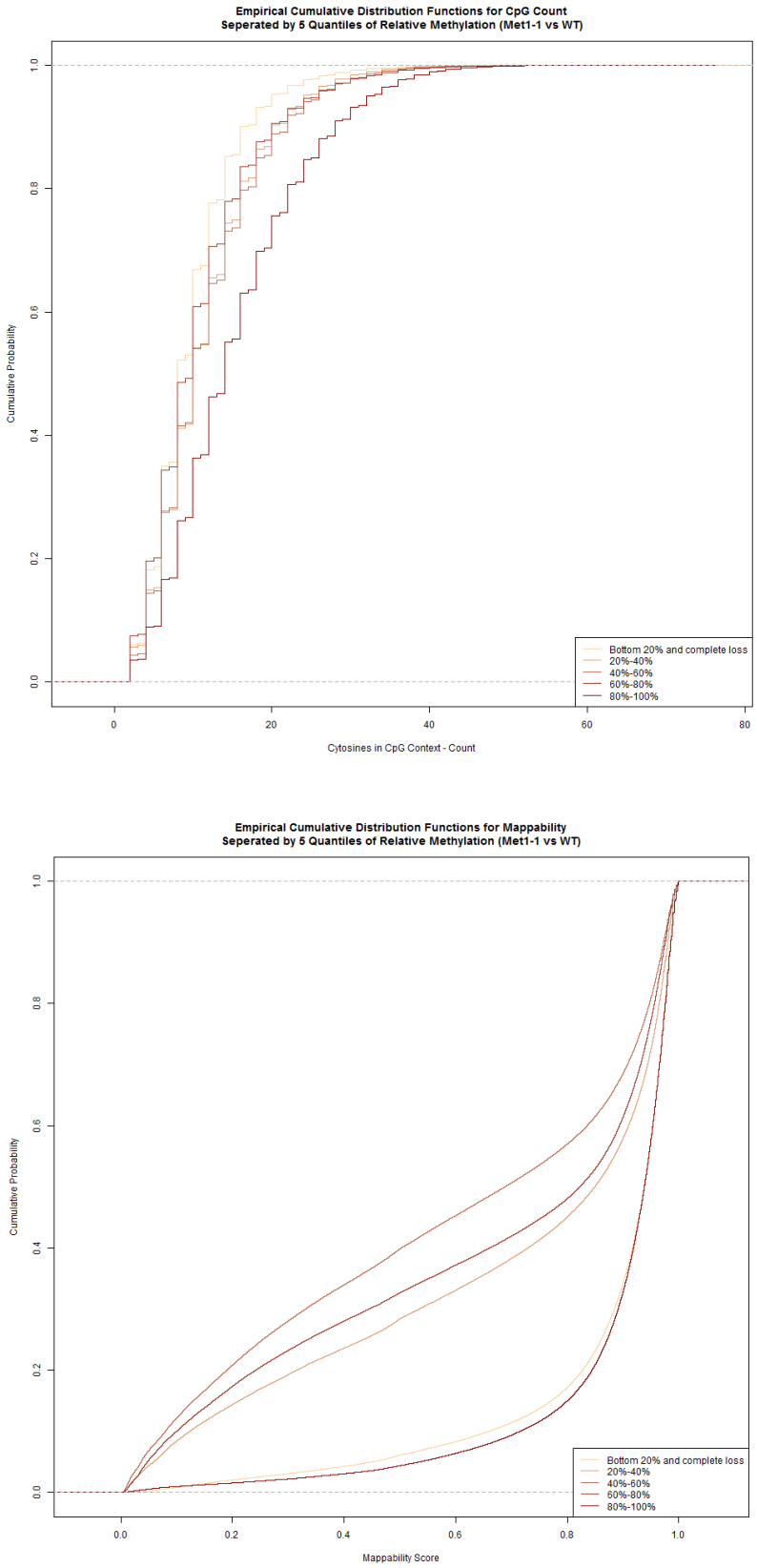

Figure 7: Empirical cumulative distribution functions for CpG context cytosine count (top) and mappability (bottom) stratified by 5 quantiles of relative methylation. Here it is clear that there is a bigger difference in the distributions of mappability than for CpG count.

From figure 7 and tables 3 & 4, it is clear that neither CpG context cytosine count nor mappability score alone is sufficient to fully distinguish the quantiles of relative methylation from one another since the distributions overlap in many cases. As an example, for mappability the top (80%-100%) and bottom (0%-20%) are incredibly similar with a Kolmogorov-Smirnov test statistic of just 0.023.

Discretised quantile model

In order to incorporate the information from both variables into a predictive classifier a first model was built by producing a joint probability distribution. In order to evaluate such a model, before building a more mathematically involved model, a simple quantile based grouping model was developed. First the tiles were assigned their actual relative methylation quantile based upon the full data set. Five quantiles were used to compare with the above analysis. Next the data was split into a training set of 70% of the data, with the remainder forming a test set. Relative methylation, CpG context cytosine count and mappability were split into 5 quantiles based on the training set. Each tile was then classed into one of the five quantiles for each variable using these breaks, effectively discretising these variables. This was applied to both sets of data.

Once the tiles were given these discrete scores, the training set data was aggregated according to the quantiles for CpG count and mappability, to create 25 possible combinations of these variables. Within each combination, the total number of tiles in each quantile of relative methylation were calculated. Using equation 1 below, these totals were then used to calculate conditional probabilities that a given tile, with known

CpG count and mappability quantiles, belonged to a given relative methylation quantile.

$$P(Rel\ Meth|CpG, Map) = \frac{P(Rel\ Meth, CpG, Map)}{\sum_{rel\ Meth} P(Rel\ Meth, CpG, Map)} \quad (1)$$

Here each variable takes one of five values, the quantiles, and the probabilities on the right-hand side are calculated by normalising the counts for each relative methylation quantile within each combination of the other variables by the total number of tiles. The summation is over all possible relative methylation quantiles for a given combination of CpG count and mappability.

The model was then used to make predictions for the test set. For each tile the quantiles for CpG count and mappability were used to calculate the appropriate probabilities. The tile was then assigned a predicted relative methylation quantile based on these probabilities by picking the quantile with the greatest probability. The results for this model used on the test set are summarised in table 5. In particular, there is a 37.3% accuracy on the test set. As the quantiles are approximately equal, this is a marked improvement on randomly placing each tile, where an accuracy of 20% would be expected.

Accuracy	Recall Top Class	Precision Top Class	F-Score Top Class	Recall Bottom Class	Precision Bottom Class	F-Score Bottom Class
37.32%	40.75%	39.33%	0.400	63.93%	37.53%	0.473

Table 5: Result for the quantile-based classifier. Recall measures the proportion of tiles actually in each class that are predicted as in that class. Precision measures the number of correct predictions for each class as a proportion of all tiles predicted to be in that class. The F-score is a combined measure of these two performance metrics (see Materials and methods)

Fitted distributions model

To build on the limited success of the joint probability density model, a model was built where the CpG context cytosine count and mappability variables were fitted to probability density functions. These could then be combined using Bayesian statistics to give a more complex function to produce probabilities for each quantile of relative methylation. This enables a better resolution, based on the scores of these variables, than a very coarse quantile discretisation approach allows.

Each tile's CpG count is necessarily a discrete variable by nature of it being a count. Hence, in attempting to fit a known probability distribution, the choice was made from those for discrete variables. Once separated by quantiles of relative methylation, the CpG context cytosine count appears to follow a negative binomial distribution. Figure 8 shows the distributions of CpG count for five quantiles of relative methylation across all data, with fitted negative binomial distributions in each case. Given the close fit of these negative binomials, it was decided to use such distributions in the final probability model.

As figure 9 shows, the distributions of mappability scores are less easily able to be fitted to a known probability distribution. As such, the probability density function for the mappability distributions within each quantile of relative methylation were estimated using a Gaussian kernel approximation method (see Materials and Methods).

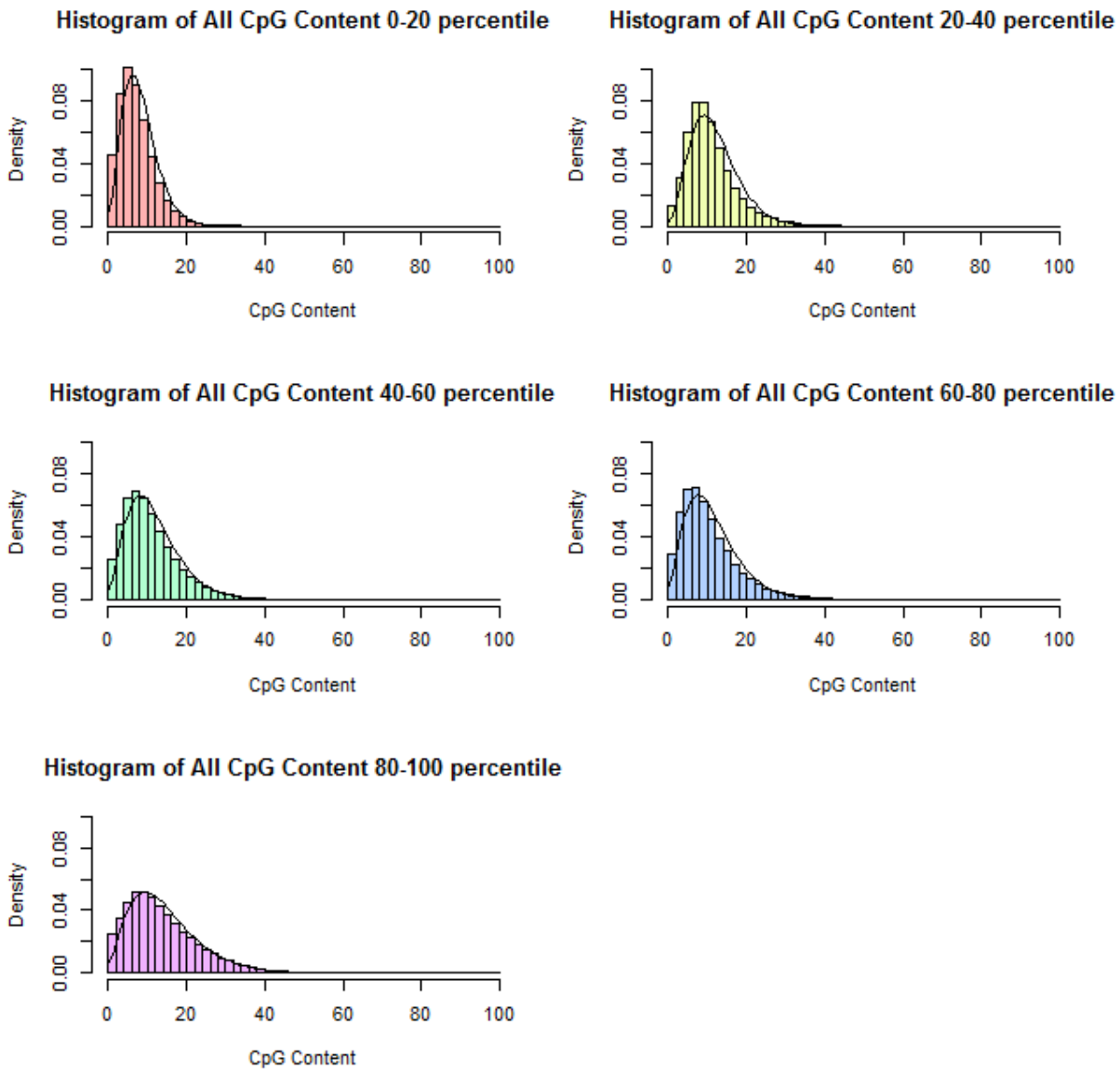

Figure 8: Histograms of CpG context cytosine count for tiles stratified by quantiles of relative methylation, with fitted binomial distributions overlaid. There is a clear trend for both an increase in CpG content and for an increase in spread of values for tiles with higher relative methylation. The fitted negative binomials show a good approximation to the distributions.

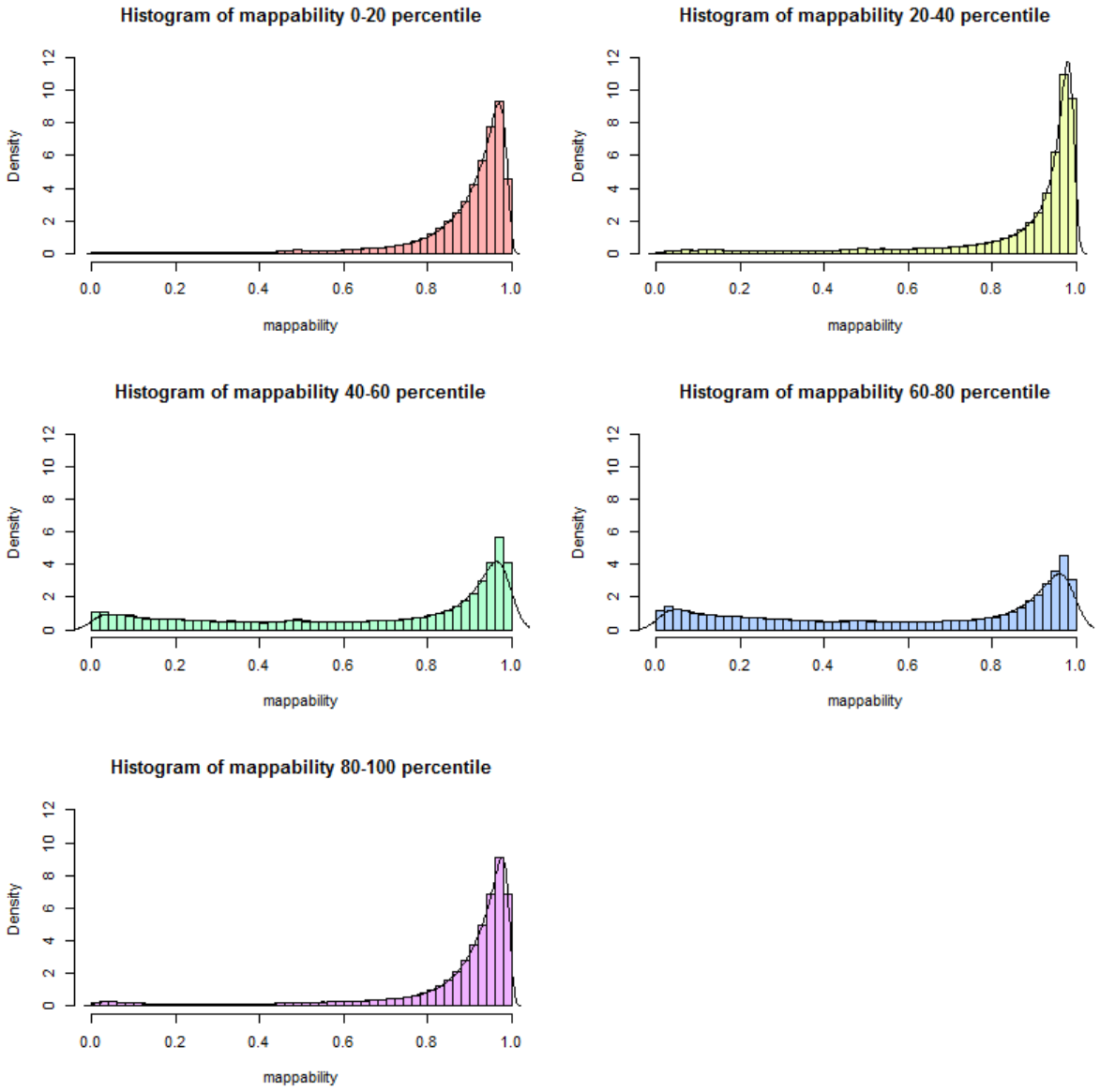

Figure 9: Histograms of mappability for tiles stratified by quantiles of relative methylation, with Gaussian kernel approximated density function. It is possible to see the increased number of tiles with low mappability in the middle quantiles as seen in figure 6. Again, the fitted distributions are a close approximation to the underlying data.

In order to combine these two sets of distributions to create a function that calculates the probability of a given tile belonging to each quantile, a Bayesian approach was once again utilised. Here, for simplicity, the distributions for CpG context cytosine count and mappability were considered to be independent. The probabilities that a given tile belonged to each quantile were then computed using equation 2. In this equation, the two probabilities in the numerator on the right-hand side are exactly the distributions given previously, and the summation in the denominator is of the product of these for each given quantile.

$$P(Rel\ Meth|CpG, Map) = \frac{P(Rel\ Meth|CpG)P(Rel\ Meth|Map)}{\sum_{rel\ Meth} P(Rel\ Meth|CpG)P(Rel\ Meth|Map)} \quad (2)$$

In order to compare the fitted distribution model to the discretised quantile model, this classifier was built with 5 quantiles for relative methylation. The model was trained on 70% of the data set, and then the probabilities of each tile belonging to each quantile of relative methylation were computed across the complete data set. Each tile was then assigned a predicted quantile by selecting the quantile with highest probability. The results for this classifier when scored on the 30% test set of the data are summarised in table 6.

Model	Accuracy	Recall Top Class	Precision Top Class	F-Score Top Class	Recall Bottom Class	Precision Bottom Class	F-Score Bottom Class
Discrete Quantiles	37.32%	40.75%	39.33%	0.400	63.93%	37.53%	0.473
Fitted Distributions	37.52%	35.36%	41.6%	0.383	66.84%	37.54%	0.481

Table 6: Comparison of classifier performance for discrete quantile and fitted distribution models. The Fitted distribution model is marginally more accurate. It scores slightly worse on the top class, but does better on the bottom class.

Both classifiers appear to score relatively similarly, with a slight improvement in accuracy for the fitted distributions model. This model does score noticeably better on

the lowest quantile however. This quantile primarily contains all tiles which have total loss of methylation. As the motivation for this predictive model was to be able to predict an ordering of tiles based on how readily they regain methylation (measured by the proxy of relative methylation under *met*1-1 mutation) it was decided to run the model again, this time filtering out all tiles which have complete methylation loss. Thus only tiles with some methylation under the mutation were used.

When running this model, it was decided to run a cross-validation regime. In this case, the data was split into a training set, a cross-validation set and a test set in the ratio 60:20:20. The model was then built 8 times, with each run using a different number of quantiles, ranging from 3-10. The distributions were trained on the training set. Each classifier was then scored on its performance on the cross-validation set. The results from this are summarised in table 7.

Quantiles	Accuracy	Expected Accuracy from Random Guessing	Fold-increase in Accuracy
3	48.65%	33.33%	1.460
4	39.62%	25%	1.583
5	34.13%	20%	1.707
6	29.46%	16.67%	1.767
7	25.17%	14.29%	1.761
8	22.40%	12.5%	1.792
9	20.13%	11.11%	1.812
10	18.57%	10%	1.857

Table 7: Results of cross-validation for fitted distribution model applied to tiles without total loss of methylation. The fold-increase in accuracy generally increases with number of classes.

Table 7 highlights that the model consistently performs better than random guessing, however it struggles to achieve high accuracy. Noting that the fold-increase in accuracy begins to increase more slowly beyond 5 classes, it was decided that continuing to use 5 classes for the model was an appropriate choice. On the test set this model scored an accuracy of 33.99%. This is in line with the score in the cross-validation step. It is also worth noting that this score is lower than when tiles with total loss of methylation are included. Hence, the model taken forward was chosen to be a 5 quantile model including complete loss tiles.

Stratification by annotation

As the motivation behind developing this model was the differing behaviour witnessed between different elements of the genome it was decided to extend the model by stratifying it based upon the genome annotation. For this purpose, a simplified annotation was created for the A*. thaliana* genome. This took an existing annotation (see Materials and Methods) and simplified each nucleotide to be labelled as either an exon, intron or intergenic. This was then overlaid with a list of nucleotides annotated as transposable elements from a public annotation. Where these overlapped exons and introns, the transposable element annotation took priority. Figure 10 shows this process. This annotation was then applied to the tiles in the data set. Where a tile overlapped more than one feature it was listed as having a non-unique annotation feature.

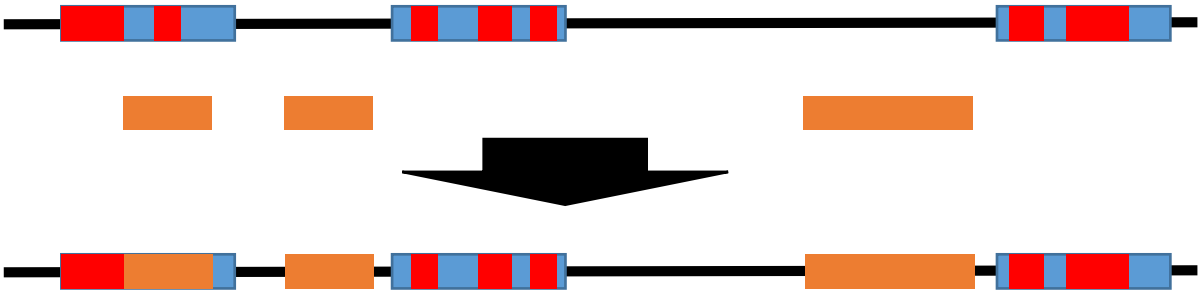

Figure 10: Construction of a simplified annotation. A reduced annotation was first constructed from a database, with exons (red), introns (blue) and intergenic (black) regions annotated. A list of candidate transposable elements (orange) was then superimposed onto this skeleton annotation. The hierarchy of annotation was transposable elements, exons, introns, intergenic regions, such that any nucleotide with multiple annotations was listed under the first to appear in this list.

The tiles were stratified by their annotation and the 5 quantile model in the previous section was applied to each annotation group, using a 70:30 training and test split. The results of these classifiers are summarised in table 8. As becomes clear, stratification does not appear to improve results across all types of annotated element. In particular, results for exons and transposable elements actually worsen. Hence such a model is less useful in terms of dealing with predictions relating to candidate TEL and GEL tiles.

Annotation	Accuracy	Recall Top Class	Precision Top Class	F-Score Top Class	Recall Bottom Class	Precision Bottom Class	F-Score Bottom Class
Transposable Element	29.15%	51.28%	28.60%	0.367	50.63%	33.11%	0.400
Exon	31.97%	52.25%	33.31%	0.407	53.52%	32.97%	0.408
Intron	44.96%	44.26%	21.76%	0.292	61.62%	68.94%	0.651
Intergenic	36.32%	41.45%	31.57%	0.358	63.50%	46.01%	0.534
Non-unique	35.59%	37.04%	39.09%	0.380	58.05%	42.98%	0.494
Combined	37.52%	35.36%	41.6%	0.383	66.84%	37.54%	0.481

Table 8: Results using a 5 class fitted function model once stratified by annotation. The results are mixed, with accuracy reducing in most cases. Intron results are biased owing to a large number of complete loss tiles causing unequal sizes for the classes.

Discussion

The models presented here struggle to produce great accuracy in predicting the relative methylation of regions of the *A. thaliana* genome under the *met*1-1 mutation. This suggests that there is likely to be more factors involved than simply CpG context cytosine content and repetitiveness of each region. The models do however show a marked improvement on random guessing. This suggests that these two factors are able to influence the differential behaviour of regions of the genome under reduced methylation conditions. This agree with the earlier work on the roles of these factors (Mathieu et al., 2007; Teixeira et al., 2009). Stratification by genomic feature failed to improve the model, which strengthens the argument that the mechanism for return of methylation is independent of the genomic feature, possibly depending only upon the underlying sequence of each region. Additionally, methylation behaviour has been observed to be similar across many species of plant and so any mechanism supported by this model may generalise to other species (Feng et al., 2010).

There are major considerations to be made for the model's limitations. Firstly, the use of methylation levels under the *met*1-1 mutation as a proxy for transgenerational effects is unlikely to truly capture all of this behaviour. With more time, this variable could be replaced with a better measure of relative rate of methylation return. The Paszkowski group have data for complete loss of MET1 function mutants that have been crossed with wild-type individuals across many generations. This data was not used here as it would first be necessary to ascertain which regions of the genome come from the loss of function epialleles as homologous recombination will cause

these to become mixed with wild-type alleles. This is a lengthy undertaking, but is certainly a good future avenue.

Another possible hindrance to any model of this process is the documented phenomenon of paramutation (Erhard & Hollick, 2011). In this process, alleles are able to act in trans to silence each other. This could lead to dilution of epiallele strength in crosses as the demethylated epiallele is targeted via such a mechanism by the wild-type allele. This process is something that could be factored in to future models.

Clearly, the models as presented here aren't sufficiently accurate to be considered predictive models. They do however support the idea that CpG content and repetitiveness are involved in the process of remethylation. However, either further factors are needed to strengthen the model or simply there is too much stochasticity in the system for this model to explain the full behaviour. The model could be extended by exploring if the size of the tiles used in the model have an effect on the power of prediction. Looking at smaller tiles may help given the involvement of small RNAs in the de novo methylation process (Law & Jacobsen, 2010). Should a more powerful predictive model be built it would also be good verification to insert artificial sections of DNA into an *A. thaliana* genome and see if the methylation behave is as the model predicts.

Materials and Methods

Bisulphite Sequencing

Though not completed as part of the project, the data used came from bisulphite sequencing experiments for various replicates of plants grown by the Paszkowski group. This method uses bisulphite reactions, causing unmethylated cytosines to convert to thymines without affecting those which are methylated. This enables sequencing to measure which cytosines were methylated in the sample (Cokus et al., 2008)

BLAST

The data presented for the project included BLAST scores presented as the number of hits scored by BLAST when run on each tile against the whole *A. thaliana* genome.

Mappability

The raw mappability data was prepared by Dr. Catoni using a local genome browser and presented as a bedgraph file. The analysis of the data was then conducted in R.

R

All routines were run in the R environment on a Windows computer with 64-bit operating system. Version 3.2.3 of R was used. All code used for this report is available at https://github.com/jonvw28/arabidopsis_cytosine_methylation

K-means & Fuzzy C-means clustering

These algorithms were applied by using the built in kmeans function in R and the cmeans function from the package “e1071” respectively. The fuzzy c-means algorithm gives weights for each data point of belonging to each cluster, as opposed to classing each point to one cluster only (Ghosh & Dubey, 2013).

Kolmogorov-Smirnov tests and Empirical Cumulative Distribution Functions (ecdf)

The ecdfs used in this report were created using the inbuilt function ecdf in R. They give an estimate of the cumulative distribution of data weighting each value by the number of times it occurs. Kolmogorov-Smirnov tests then compare these, scoring the statistical significance of the greatest difference between these. This was run using the ks.test function in R (Arnold & Emerson, 2011).

Kernel Estimation

This is a method to estimate a distribution based upon weighted kernels. In the report, Gaussian kernels were used, using the built in R function density with default settings.

F-Score

This is a measure of classifier performance based upon its precision and recall. It is calculated as $\frac{2\,Precision \times Recall}{Precision + Recall}$

Annotations

All annotations were taken from the Arabidopsis Information Resource (TAIR) and can be found at https://www.arabidopsis.org/

Acknowledgements

I would like to thank Marco Catoni for providing me with the idea behind this project, for providing me with the raw data used to complete it and for supporting me throughout.

I would also like to thank Hajk Drost for his support and guidance offered on the computational side of this project, and for his advice in developing and applying many of the ideas underling my models.

Finally, I would like to thank Jerzy Paszkowski and the rest of the Paszkowski group for their advice and feedback throughout the project. Their opinions helped to shape the path this project took, and as a group they were very supportive and welcoming.

References

Arnold, T., & Emerson, J. (2011). Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions. *The R Journal*, 34–39. Retrieved from http://journal.r-project.org/archive/2011-2/RJournal_2011-2_Arnold+Emerson.pdf

Bartee, L., Malagnac, F., & Bender, J. (2001). Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. *Genes and Development*, *15*(14), 1753–1758. http://doi.org/10.1101/gad.905701

Calarco, J. P., Borges, F., Donoghue, M. T. A., Van Ex, F., Jullien, P. E., Lopes, T., … Martienssen, R. A. (2012). Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell*, *151*(1), 194–205. http://doi.org/10.1016/j.cell.2012.09.001

Chan, S. W.-L., Henderson, I. R., & Jacobsen, S. E. (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nature Reviews. Genetics*, *6*(5), 351–360. http://doi.org/10.1038/nrg1664

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., … Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, *452*(7184), 215–219. http://doi.org/10.1038/nature06745

Erhard, K. F., & Hollick, J. B. (2011). Paramutation: A process for acquiring trans-generational regulatory states. *Current Opinion in Plant Biology*, *14*(2), 210–216. http://doi.org/10.1016/j.pbi.2011.02.005

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., … Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(19), 8689–8694. http://doi.org/10.1073/pnas.1002720107

Ghosh, S., & Dubey, S. K. S. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. *Ijacsa*, *4*(4), 35–38. http://doi.org/10.14569/IJACSA.2013.040406

Goll, M. G., & Bestor, T. H. (2005). Eukaryotic Cytosine Methyltransferases. *Annual Review of Biochemistry*, *74*(1), 481–514. http://doi.org/10.1146/annurev.biochem.74.010904.153721

Heard, E., & Martienssen, R. A. (2014). Transgenerational epigenetic inheritance: Myths and mechanisms. *Cell*, *157*(1), 95–109. http://doi.org/10.1016/j.cell.2014.02.045

Johnson, L. M., Bostick, M., Zhang, X., Kraft, E., Henderson, I., Callis, J., & Jacobsen, S. E. (2007). The SRA Methyl-Cytosine-Binding Domain Links DNA and Histone Methylation. *Current Biology*, *17*(4), 379–384. http://doi.org/10.1016/j.cub.2007.01.009

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews. Genetics*, *13*(7), 484–92. http://doi.org/10.1038/nrg3230

Kakutani, T., Jeddeloh, J. a, Flowers, S. K., Munakata, K., & Richards, E. J. (1996). Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(22), 12406–12411. http://doi.org/10.1073/pnas.93.22.12406

Kakutani, T., Munakata, K., Richards, E. J., & Hirochika, H. (1999). Meiotically and mitotically stable inheritance of DNA hypomethylation induced by ddm1 mutation of Arabidopsis thaliana. *Genetics*, *151*(2), 831–838.

Kankel, M. W., Ramsey, D. E., Stokes, T. L., Flowers, S. K., Haag, J. R., Jeddeloh, J. A., … Richards, E. J. (2003). Arabidopsis MET1 cytosine methyltransferase mutants. *Genetics*, *163*(3), 1109–1122.

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews. Genetics*, *11*(3), 204–220. http://doi.org/10.1038/nrg2719

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. http://doi.org/10.1016/j.cell.2008.03.029

Mathieu, O., Reinders, J., Čaikovski, M., Smathajitt, C., & Paszkowski, J. (2007). Transgenerational Stability of the Arabidopsis Epigenome Is Coordinated by CG Methylation. *Cell*, *130*(5), 851–862. http://doi.org/10.1016/j.cell.2007.07.007

Matzke, M. A., & Mosher, R. A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature Reviews Genetics*, *15*(6), 394–408. http://doi.org/10.1038/nrg3683

Matzke, M., Kanno, T., Daxinger, L., Huettel, B., & Matzke, A. J. (2009). RNA-mediated chromatin-based silencing in plants. *Current Opinion in Cell Biology*, *21*(3), 367–376. http://doi.org/10.1016/j.ceb.2009.01.025

Mirouze, M., & Paszkowski, J. (2011). Epigenetic contribution to stress adaptation in plants. *Current Opinion in Plant Biology*, *14*(3), 267–274. http://doi.org/10.1016/j.pbi.2011.03.004

Miura, a, Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H., & Kakutani, T. (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature*, *411*(6834), 212–214. http://doi.org/10.1038/35075612

Paszkowski, J., & Grossniklaus, U. (2011). Selected aspects of transgenerational epigenetic inheritance and resetting in plants. *Current Opinion in Plant Biology*, *14*(2), 195–203. http://doi.org/10.1016/j.pbi.2011.01.002

Pickford, a S., & Cogoni, C. (2003). RNA-mediated gene silencing. *Cellular and Molecular Life Sciences : CMLS*, *60*(5), 871–882. http://doi.org/10.1007/s00018-003-2245-2

Rabinowicz, P. D., Palmer, L. E., May, B. P., Hemann, M. T., Lowe, S. W., Richard, W., … Martienssen, R. A. (2003). Genes and Transposons Are Differentially Methylated in Plants , but Not in Mammals Genes and Transposons Are Differentially Methylated in Plants , but Not in Mammals, 2658–2664. http://doi.org/10.1101/gr.1784803

Saze, H., Mittelsten Scheid, O., & Paszkowski, J. (2003). Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nat Genet*, *34*(1), 65–9. http://doi.org/10.1038/ng1138

Selker, E. U. (1999). Gene Silencing. *Cell*, *97*(2), 157–160. http://doi.org/10.1016/S0092-8674(00)80725-4

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., … Jacobsen, S. E. (2014). Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nature Structural & Molecular Biology*, *21*(1), 64–72. http://doi.org/10.1038/nsmb.2735

Takuno, S., & Gaut, B. S. (2013). Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(5), 1797–802. http://doi.org/10.1073/pnas.1215380110

Teixeira, F. K., Heredia, F., Sarazin, A., Roudier, F., Boccara, M., Ciaudo, C., … Colot, V. (2009). A Role for RNAi in the Selective. *Science*, *323*(March), 1600–1604.

Tompa, R., McCallum, C. M., Delrow, J., Henikoff, J. G., Van Steensel, B., & Henikoff, S. (2002). Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. *Current Biology*, *12*(1), 65–68. http://doi.org/10.1016/S0960-9822(01)00622-4

Tran, R. K., Henikoff, J. G., Zilberman, D., Ditt, R. F., Jacobsen, S. E., & Henikoff, S. (2005). DNA Methylation Profiling Identifies CG Methylation Clusters in Arabidopsis Genes Robert. *Current Biology*, *15*, 154–159. http://doi.org/10.1016/j

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., & Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, *461*(7262), 423–426. http://doi.org/10.1038/nature08351

Wassenegger, M., Heimes, S., Riedel, L., & Sanger, H. L. (1994). RNA-directed De-Novo methylation of genomic sequences in plants. *Cell*, *76*(3), 567–576.

Weigel, D., & Colot, V. (2012). Epialleles in plant evolution. *Genome Biology*, *13*(10), 249. http://doi.org/10.1186/gb-2012-13-10-249

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W. L., Chen, H., … Ecker, J. (2006). Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell*, *126*(6), 1189–1201. http://doi.org/10.1016/j.cell.2006.08.003

Zilberman, D., Cao, X., Johansen, L. K., Xie, Z., Carrington, J. C., & Jacobsen, S. E. (2004). Role of Arabidopsis ARGONAUTE4 in RNA-Directed DNA Methylation Triggered by Inverted Repeats. *Current Biology*, *14*, 1214–1220. http://doi.org/10.1016/j

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics*, *39*(1), 61–69. http://doi.org/10.1038/ng1929