EmbryoGENE Methylation Analysis Pipeline

A manual for the analysis of hybridization results for the EmbryoGENE bovine epigenetics platform

Table of contents

1. Platform information

1.1 Platform summary

The EmbryoGENE epigenetic platform allows the study of methylation and hydroxymethylation in the bovine epigenome. It is based on an Agilent manufactured 2 x 400K oligo-array which contains a total of 414,566 probes surveying 20,355 genes and 34,379 CpG islands. All experiments using the EmbryoGENE epigenetic platform involve at least three steps: (1) genomic digestion using the MseI restriction enzyme, (2) methylation-sensitive fragment selection and (3) microarray hybridization.

To maximize coverage while reducing costs, the EmbryoGENE epigenetic slide was designed assuming that all samples are first subjected to a genomic digestion using the MseI restriction enzyme. This yields a predictable set of genomic fragments, and each probe on the microarray is designed to measure the abundance of one of those fragments.

The second step, fragment selection, can be done either through methylation-sensitive digestion and ligation mediated amplification PCR (LMA-PCR) or through the use of methyl-binding proteins (MBP). In the first of those methods, adapters are ligated to the MseI genomic fragments, which are then subjected to methyl-sensitive restriction enzymes. Unmethylated fragments are cut, and thus cannot be amplified in the following PCR (Figure 1). The EmbryoGENE bovine epigenetics platform was designed with a mix of restriction enzymes in mind (Table 1). In the second method, fragmented DNA is put in contact with magnetic beads coated with MBPs which retain methylated DNA for selective enrichment. While there are currently no standard operating protocol for this type of fragment selection, EDMA's oligos were designed in such a way to be compatible with it.

Nom	Site	Sensitivity
MseI	T/TAA	-
HpaII	C/CGG	5mC
AciI	C/CGC	5mC
HinP1I	GC/GC	5mC

Table 1. Enzymes involved in the LMA-PCR protocol.

Once fragments are selected, they can be hybridized to the EDMA microarray to determine their abundance.

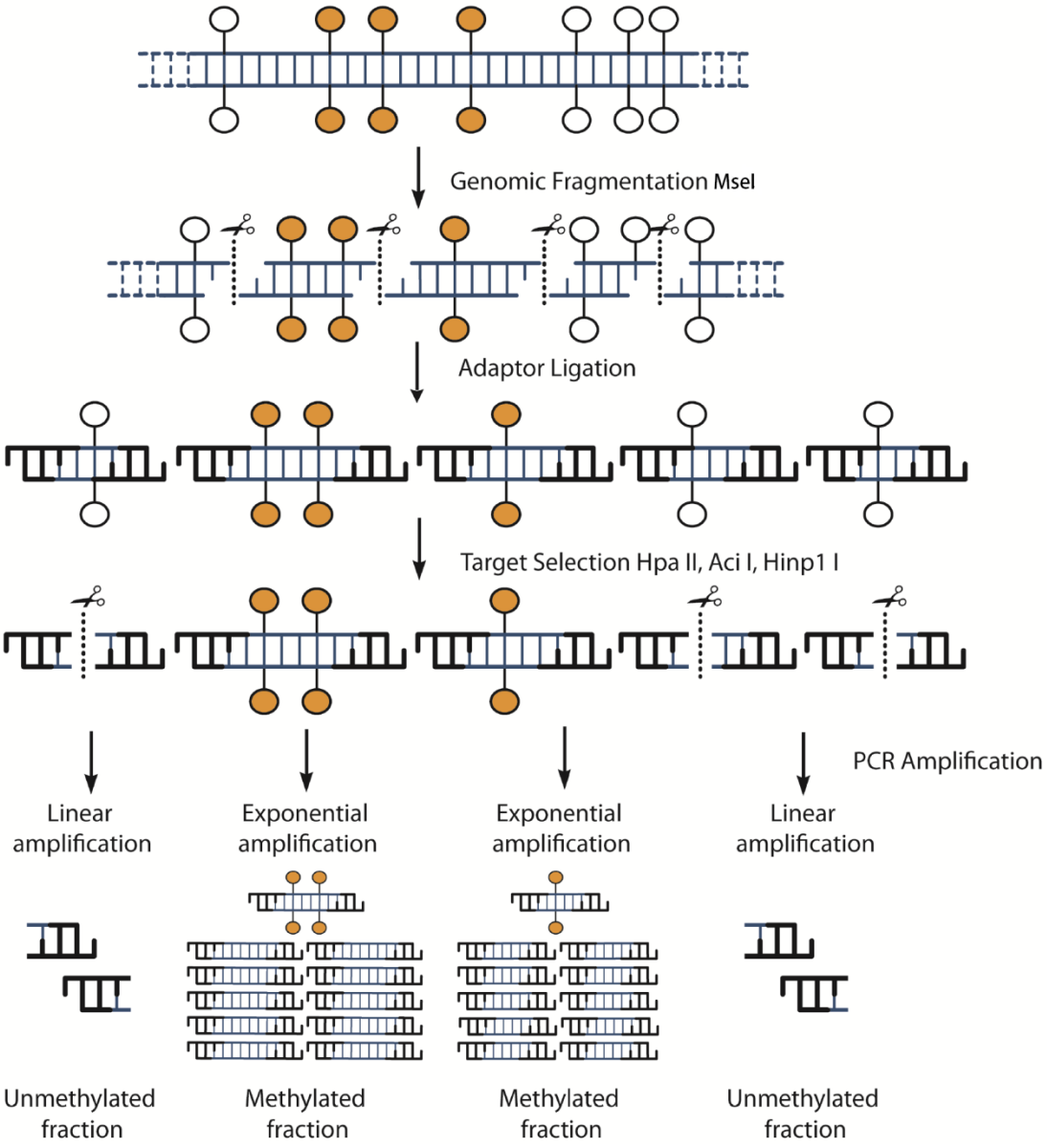

Figure 1. Fragment selection by methylation sensitive digestion and LMA-PCR. (Adapted from *De Montera, 2013*)

1.2 Probes

1.2.1 Control probes

The EDMA microarray contains three types of control probes

Agilent controls

Standard Agilent controls. The most important is (-)3xSLv1, which can be used to estimate the background intensity level across the slide. More information about Agilent controls is available in the following Agilent brochure.

Digestion controls

Digestion controls are designed with an MseI restriction site at their center and are tiled at every 1M base pairs across the entire bovine genome. They can be used to assess how complete the genomic digestion has been. Their probe names begin with EDMA_DIG.

Spike-in controls

The EDMA spike-in controls are exogenous DNA fragments which were chosen for their lack of homology to the bovine genome and for the presence of specific HpaII, AciI and HinP1I restriction sites within their sequence. They have been artificially methylated or demethylated to provide positive and negative controls for the methylation sensitive digestion step. Their probe names begin with EDMA_SPK.

1.2.2 Methylation probes

All EDMA probes which are not controls are meant to measure genomic methylation and/or hydroxymethylation. The regions selected for probe design were chosen either because they overlapped CpG islands or because they were identified in preliminary experiments which are detailed in De Montera, 2013.

1.2.3 Probe naming convention

All probe names start with "EDMA_TYP", where TYP is the probe **type**, either MET (Methylation, standard probe), DIG (Digestion control) or SPK (Spike). For MET and DIG probes, this is followed by "_XX_YYYYY", where XX is the **chromosome number** (30 for chromosome X) and YYYYY is the **sequence number of the probe** on the chromosome, ie, the first probe on the chromosome is labeled 00001, the second 00002, etc. For SPK probes, the type is followed by _ENZ_Pos/Neg_X, where ENZ is either Hpa (for HpaII), Aci (For AciI) or Hin (for HinP1I), Pos/Neg is the control type (positive or negative) and X is a unique identifier.

1.4 Probe annotation

The full annotation for the EDMA microarray is available either as a tab-separated text file or an Excel xlsx file. The meaning of each field of the annotation follows:

Probe

Probe ID.

Sequence

Probe sequence.

Number of hits

The estimated number of genomic intervals this probe should hybridize with. For probes targeting specific intervals, this should be 1. For probes targeting repeated elements, this number may vary.

Chr / Probe Start / Probe End

For probes targeting a single interval, the position of the probe's alignment to the genome. If this probe may target multiple regions, these fields are empty.

Fragment Start/ Fragment End

For probes targeting a single interval, the position of the start and end of the MseI-MseI fragment it should hybridize to.

CpG

Number of CpG dinucleotides within the MseI-MseI fragment.

HpaII/ AciI/ HinP1I/ FspBI/ MseI

Number of restriction sites for each enzyme which can be found within the MseI-MseI fragment.

Cpg Island

Number of base pairs annotated as being part of a CpG Island within the MseI-MseI fragment.

Exon / Intron / Proximal Promoter / Promoter / Distal Promoter

These fields detail which genes/transcripts can be found within the MseI-MseI fragment targeted by the probe. For exons and introns, the format of each entry is [Gene Symbol]-[Exon/Intron Number], whereas for promoter elements, only the [Gene Symbol] is present. Multiple genes/introns/exons may be present within each field, and are separated by spaces.

The "Proximal Promoter", "Promoter" and "Distal Promoter" regions are defined as the first 1kbp , 5kbp and 50kbp 5' of the transcription start site.

Fragment/Probe Repeat Family

Name of the repeat families which are found within the sequence of the probe or of the MSeI-MseI fragment. Multiple entries are separated by spaces.

Fragment/Probe Repeat Percent

Percentage of base-pairs within the probe/MseI-MseI fragment that have been identified as being part of a repeated element.

Fragment/Probe GC Percent

GC percent within the probe/MseI-MseI fragment.

UCSC_CpG_Proximity

How close to a CpG Island the MseI-MseI fragment is. Possible values are "Open Sea" (>4kbp), "Shelf" (4kbp-2kbp), "Shore" (2kbp-1) and "Island" (A CpG island lies within the bounds of the fragment).

CpG_Length

The length of the CpG Island (if any) of which the MseI-MseI fragment targeted by this probe is part.

CpG_Density

Percent of CpG dinucleotides within the CpG island which overlaps the MseI-MseI fragment targeted by this probe.

Gene_Distance

The distance (in bp) of the closest gene from the MseI-MseI fragment. Negative values indicate genes that are upstream using the standard genomic orientation of the UMD3.1 bovine genome assembly.

EMBV3_Probe

The EMBV3 probe associated with the gene closest to the EDMA probe, if such a gene exists within 50kbp of the MseI-MseI fragment.

2. Analysis

2.1 Image file analysis

Image files must be converted to intensity files before they can be analyzed. This process can be completed using the ArrayPro software. EmbryoGENE has produced a guide on using ArrayPro to analyze microarray scans (french).

2.2 Intensity file analysis

2.2.1 ELMA

The EmbryoGENE Material Transfer Agreement states that the intensity files of all microarray hybridizations using EmbryoGENE's platforms should be deposited into the ELMA LIMS. ELMA is a MIAME compliant LIMS that was developed specifically for EmbryoGENE sponsored projects. It provides storage for both data and metadata as well as basic data analysis. To obtain an ELMA account, contact one of ELMA's administrators.

2.2.2 EMAP

The **E**mbryoGENE **M**ethylation **A**nalysis **P**ipeline is a set of R scripts specifically for the analysis of EDMA slides. Users who feel comfortable with the R environment can download and run the scripts themselves. Users who feel less adventurous can ask one of EmbryoGENE's bioinformaticians with assistance in getting the scripts running. EMAP is recommended as a first step to all EMDA microarray analysis, and sections 3, 4 and 5 deal with the output of this analysis pipeline.

2.2.2 Other analysis options

EDMA microarrays can also be analyzed through any standard microarray analysis pipeline or software. One of those software is FlexArray, whose two color component was developed in collaboration with EmbryoGENE. Training material for FlexArray produced by Genome Québec can be found here. Additionally, EmbryoGENE has produced a user guide for two-color analysis using FlexArray. A list of other microarray analysis software can be found on the EmbyroGENE genome browser's Tool page.

3. Quality control

Both ELMA and the EmbryoGENE R scripts produce two set of quality control plots, which are detailed in this section. All quality control plots are based on raw intensity data.

3.1 Digestion control analysis

Multiple probes on the array are designed to have an MseI restriction site in the middle of their sequence. They can be used to assess how complete the genomic digestion step has been. The produced boxplot (Figure 2) displays how these probes behave on a per-chromosome basis. The dashed horizontal line represents the "detection cutoff" intensity value (See section 4.1). Lower digestion control intensities, optimally below the intensity cutoff line, indicate that the genomic digestion was successful.

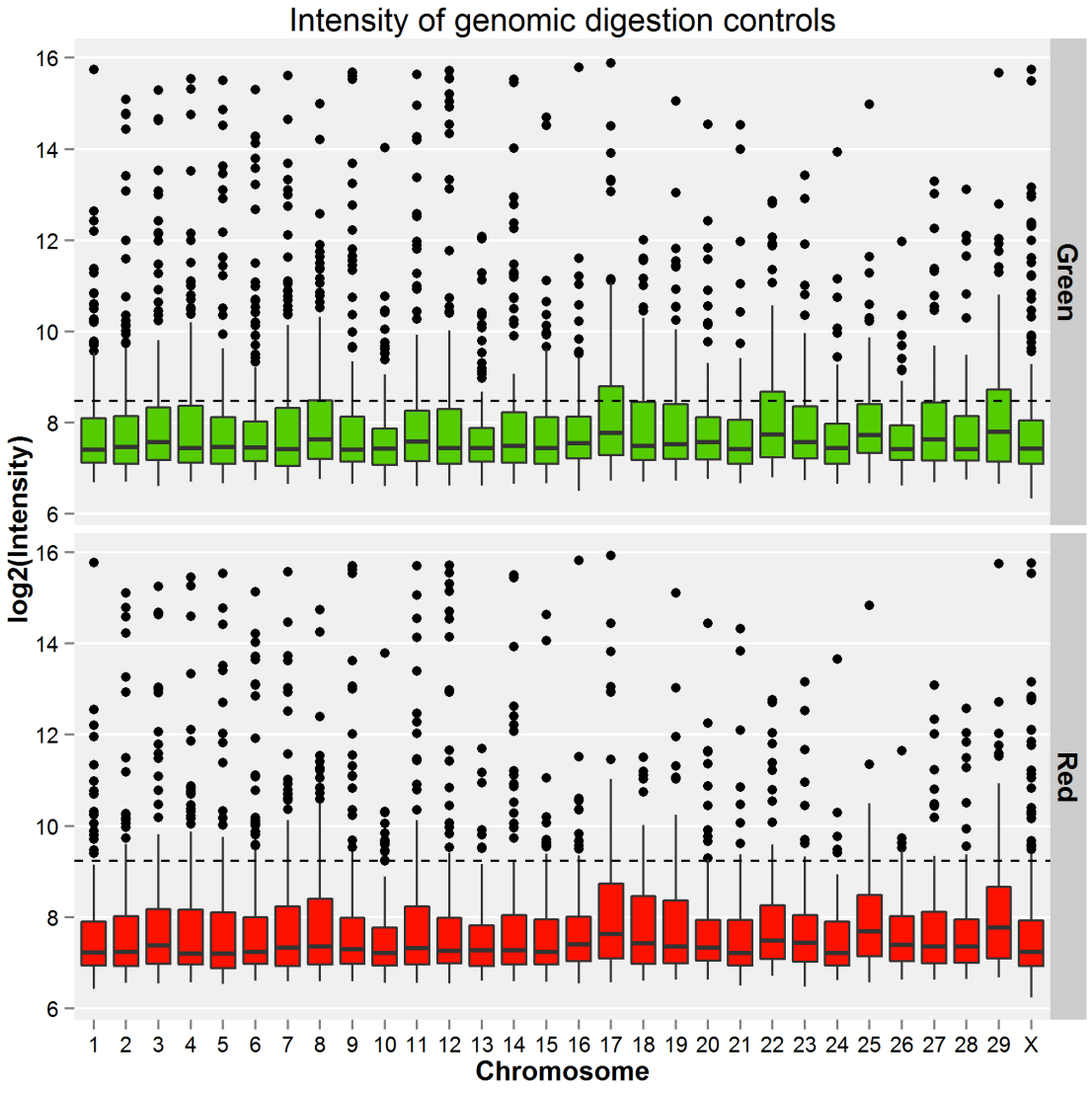

Figure 2. Example digestion control box plot.

3.1 Spike analysis

To determine the processivity of the methyl-sensitive enzymes used in the digestion step, we produced exogenous spike-in controls from tomato DNA. Those controls contain HpaII, AciI and HinP1I restriction sites which have been either artificially methylated or demethylated using methylases and PCR, respectively. The digestion of spike controls (Figure 3) plot shows an estimation of the percentage of digestion of each type of spikes as inferred from the range of intensity values of the associated probes. Lower values for non-methylated controls are best, and high values for positive controls are better.

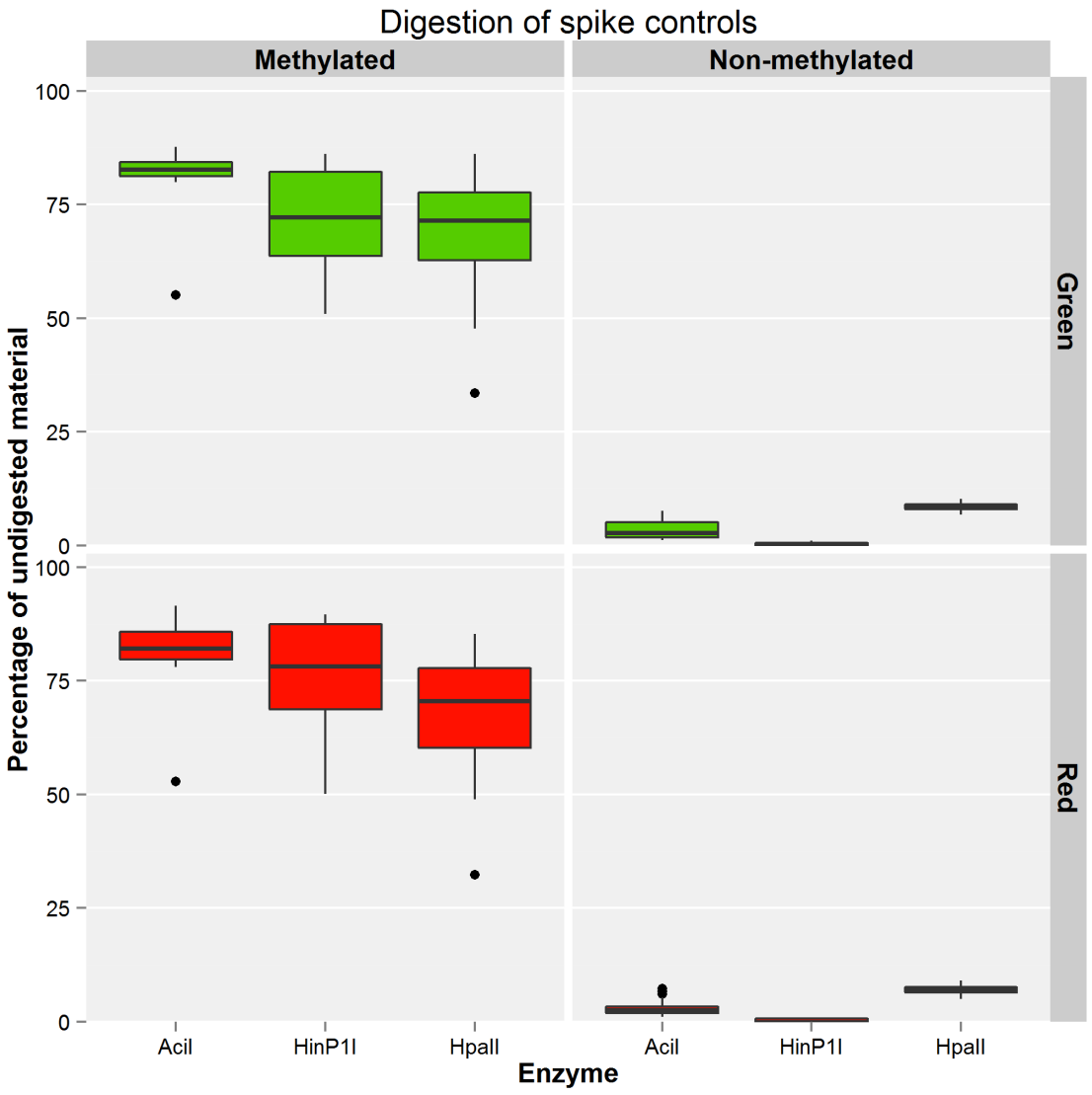

Figure 3. Example spike control box plot.

4. Limma analysis and differentially methylated region detection

Most of the statistical analysis within EMAP is performed using the Limma bioconductor package (Smyth, GK (2005)). The steps involved in this linear analysis (along with their results) are presented in this section.

4.1 Identification of probes above the background level

The EDMA microarray contains non-specific control probes ("(-)3xSLv1"), which consist of random 60-mers with no known homology to the bovine genome. These controls can be used to assess the level of "noise" on the microarray, which includes background fluorescence and non-specific binding of short nucleotide sequences. For each array, we define our **D**etection **C**utoff for array **i**, $\mathbf{DC_i}$, as:

$$DC_i = mean(NC_i) + 4 * standard\ deviation(NC_i)$$

where $\mathbf{NC_i}$ is the set of measured fluorescence intensities of all **N**egative **C**ontrol probes on array **i**. Assuming a normal distribution for background intensity, this detection cutoff should discriminate between actual signal and background noise in 99.99% of cases.

WARNING: While the detection cutoff can successfully and reliably differentiate between background noise and actual signal, having a signal does not necessarily imply methylation in the target region. Because of the combined effect of incomplete digestions and small amounts of linear amplification of digested material, non-methylated regions of the genome might exhibit some signal above the detection cutoff. Nevertheless, the probability of methylation is a direct function of signal intensity and, on the scale of a whole array, comparison to the detection cutoff is a reliable indicator of overall methylation levels.

EMAP compares the signal intensities of all probes on an array with their respective detection cutoffs, and produce the following plots:

- An histogram of the number of arrays in which probes are found to be above the cutoff, for each conditions (Figure 4).
- A comparison of the total number of probes per array which are above the detection cutoff, for each conditions (Figure 5). The indicated p-value expresses the likelihood that the number of probes above the cutoff is different between the two conditions.
- A venn diagram showing the overlap of probes above the detection cutoffs for all arrays of the reference condition, all arrays of the second condition, and the differentially methylated regions identified by the linear fit (See section 4.4) (Figure 6)

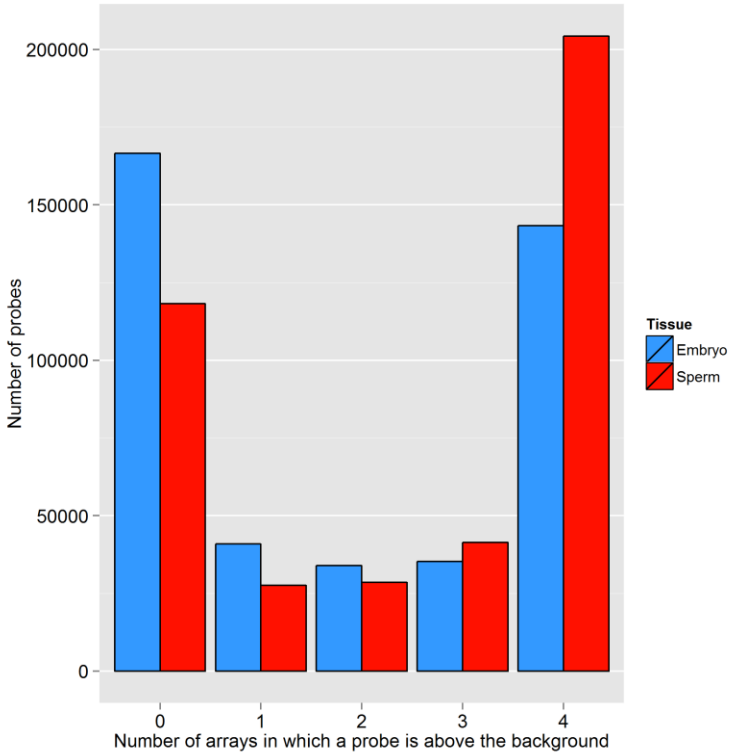

Figure 4. Histogram of the number of arrays in which probes are found to be above the detection cutoff, per condition.

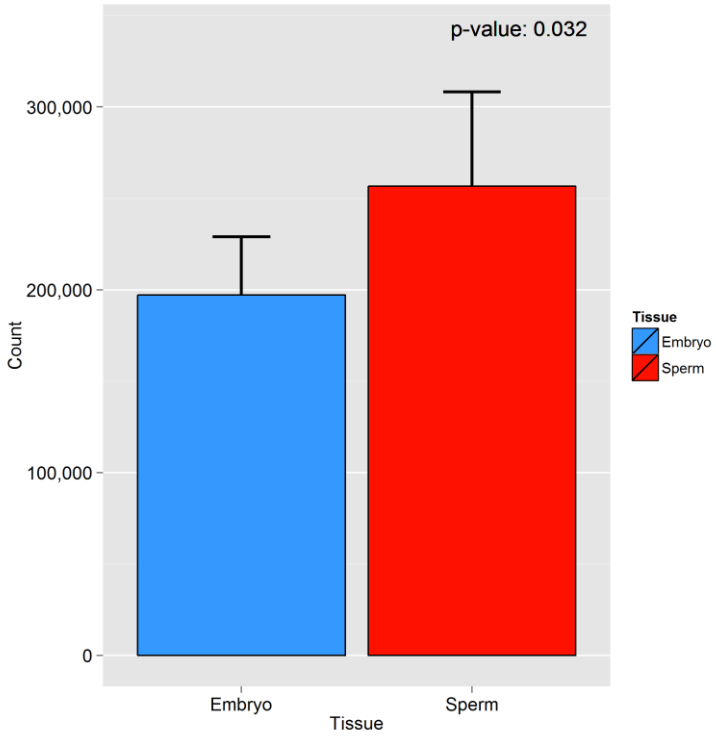

Figure 5. Total number of probes per array which are above the detection cutoff, per condition.

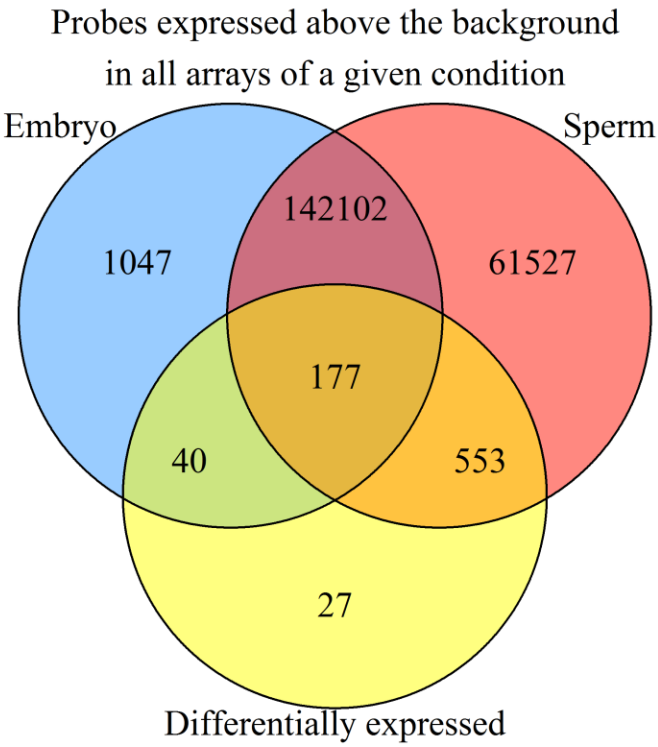

Figure 6. Venn diagram showing the overlap between probes found above the detection cutoff of all arrays in the reference condition, probes found above the detection cutoff of all arrays in the other condition, and the differentially expressed probes identified by limma.

4.2 MA Plot

Two-color microarrays are usually analyzed using M and A values. Simply put, M-values represent the $\log_2$ of the red intensity **MINUS** the $\log_2$ of the green intensity, while A-values represent the **AVERAGE** of the $\log_2$ of the intensities. Formally:

$$\boldsymbol{M} = \log_2 R - \log_2 G$$

$$\boldsymbol{A} = \frac{(\log_2 R + \log_2 G)}{2}$$

MA values are usually represented using an MA plot (Figure 7). EMAP produces an MA plot for each of the arrays in an experiment.

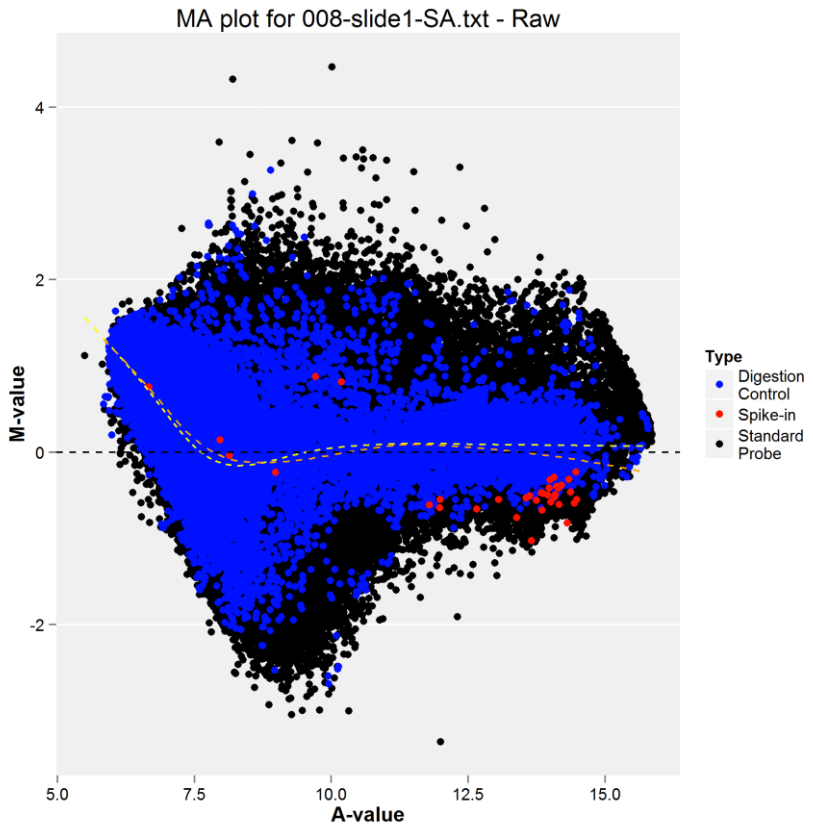

Figure 7. Example MA plot of raw data.

The EMAP MA plot contains the following elements:

- In black, the M and A values for all standard (EDMA_MET) probes.
- In blue, the M and A values for all genomic digestion controls (EDMA_DIG).
- In red, the M and A values for all exogenous spike controls (EDMA_SPK).
- The yellow dashed line represents a loess curve fitted to all of the probes on the array.
- The orange dashed line represents a weighted loess curve where half of the weight is given to the exogenous spike control, and the other half is spread amongst the genomic digestion controls. All other points are disregarded.

The MA-plot of an EDMA microarray's raw data usually shows a small bias toward positive M values at low A-values, due dye effects. Similarly, a linear bounding creating a triangular shape at low A-values is also expected from probes which show signal in only one of the two conditions.

4.3 Normalization

This raw data presented in the above MA plot is normalized using a two-step process:

1. First, a **within-array loess normalization** is applied. This process fits a loess curve to an array's MA values, then subtracts that curve from all points, leaving only the residuals.
2. Secondly, a **between array quantile normalization** which ensures that the intensities have the same empirical distribution across arrays and across channels.

The statistical details of these normalization methods are explained in Smyth, G. K., and Speed, T. P. (2003). Once the data is normalized, EMAP produces a new MA-plot (Figure 8). Normalized MA plots should be centered around 0.

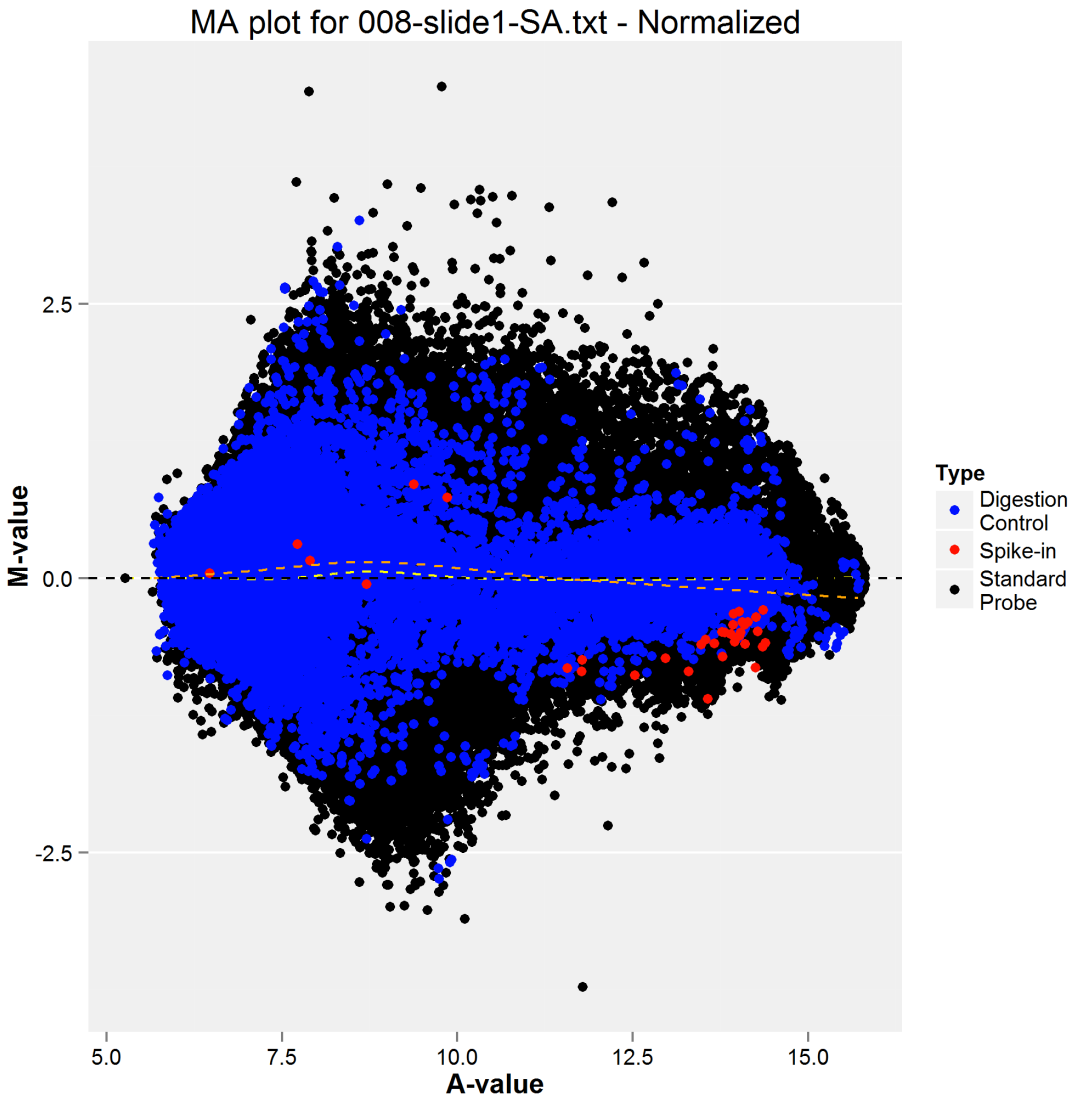

Figure 8. Example MA plot of normalized data

4.4 Linear fit

Once the data has been normalized, limma establishes fold-changes and the statistical likelihood of differential expression by fitting a gene-by-gene linear model. More details about this process can be found in chapter 8 of the Limma user guide, or within the Smyth, 2004 paper.

The net result is that each probe is associated two values: a fold-change and a p-value. The fold-change, always presented as $\log_2$(Other Condition/Reference Condition), represents the ratio of signals between the two conditions. The p-value represents the probability that the mean intensities between conditions is different. In microarray analysis, a probe is considered to be of interest if both its fold-change and p-value meet certain thresholds. By default, EMAP use a fold-change threshold of log2(1.5) and a p-value threshold of 0.05 to determine which probes constitute **D**ifferentially **M**ethylated **R**egions (DMRs).

Fold-changes and p-values are best summarized by a volcano plot, which EMAP generates for all experiments (Figure 9). In an EMAP volcano plot, the dashed lines represent the fold-change and p-value thresholds.:

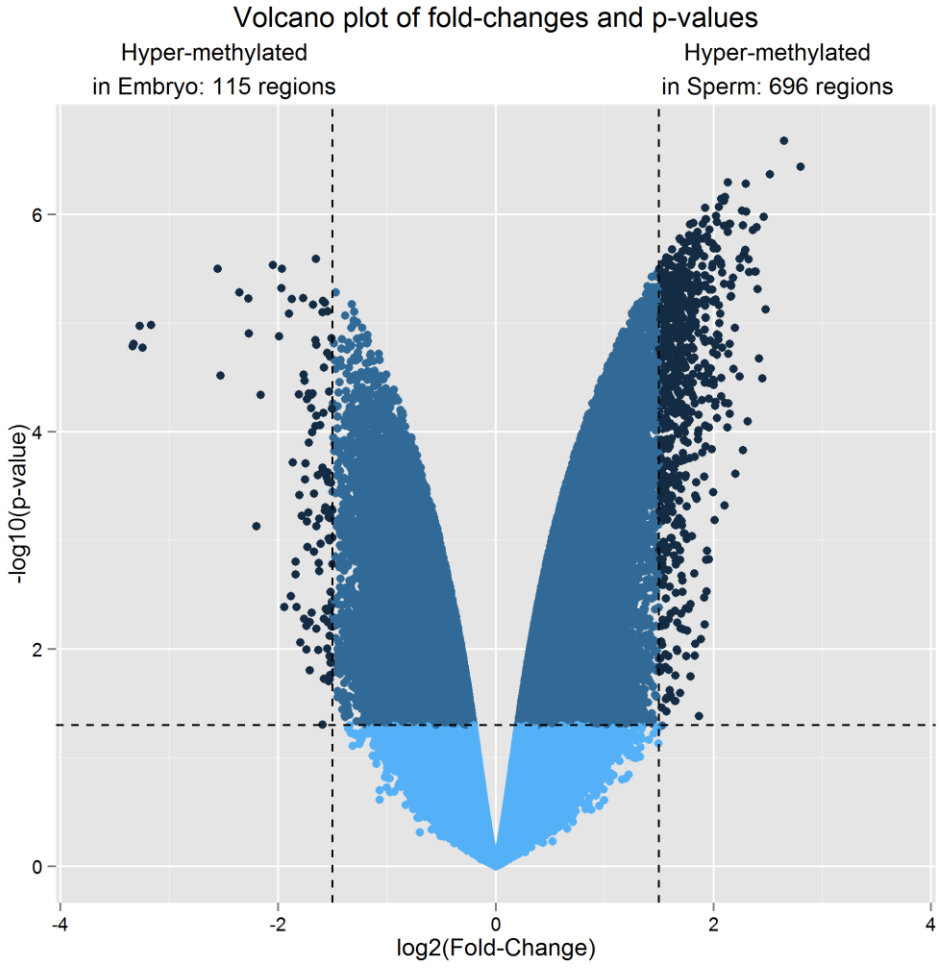

Figure 9. Volcano plot of fold-changes and p-values

The fold-changes and p-values presented in the volcano plot can be found in the LimmaAnalysis.txt text file, which is arguably the most important result file produced by the analysis pipeline. It should be opened from a spreadsheet application, such as Microsoft Excel or OpenOffice Calc, and interpreted in combination with the annotation file describing each probe.

The DiffExpr.txt file is a subset of the LimmaAnalysis.txt file containing probes which met the significance threshold established for the analysis.

WARNING: Due to the enrichment process used to select methylated genomic fragments, there is no linear, highly correlated relationship between the fold-changes calculated by limma and the level of methylation of genomic regions. Thus, while higher fold-changes imply higher odds of differential methylation, a fold-change twice as high for one gene than for another does not imply this gene is two times more methylated. Furthermore, a change in methylation in one region surrounding a gene does not always result in an equivalent change in expression of that same gene.

Thus, the best, most reliable way of interpreting the fold-changes and p-values provided by limma are as **ordered lists of probability of differential methylation**. The default threshold values have been chosen because, in most cases, they give reasonably sized DMR lists and can be used as a basis for inter-experimental comparison of the variation in methylation between two given tissues.

5. Other results

5.1 Visualisation through bedgraph files

The pipeline produces various bedgraph files which can be imported into visualization tools such as the UCSC genome browser or our own mirror at EmbryoGene. They contain an association of genomic coordinates and values. What each file represents is indicated by its the file name. Files starting with "Probe" associate values to probe coordinates, whereas files starting with "Fragment" associate the exact same values to the coordinates of the MseI-MseI restriction fragment targeted by those probes. There are bedgraph files for fold-changes, condition means and p-values. These files can be used to visualize changes around genomic regions, and are what is used to generate the circos plots which are the object of section 5.5.

5.2 Ingenuity Pathway Analysis input file

Ingenuity Pathway Analysis (IPA) requires a one-to-one association of genes and fold-changes/p-value pairs. However, since multiple probes assess methylation changes in and around most genes, and that certain probes are close to more than one gene, such a one-to-one relationship is not self-evident. EMAP solves this problem by generating the IPA.txt file, which , for all genes surveyed by the EDMA array, looks for the associated probe with the highest fold-change and lowest p-value, and associates those values with the gene's symbol. This results in an optimistic, upper-bounded estimation of methylation changes across the genome.

WARNING: Given that (as stated in section 4.4) EMAP fold-changes are not a linear function of the levels of methylation, the wisdom of using them within IPA is uncertain. The IPA.txt file is provided as a service, but each individual user should determine if such an analysis makes sense in the context of his or her experiment.

5.3 Methylation variation hotspots

The HotSpots.txt file contains averages of p-values of differential methylation over windows of 100K nucleotides. More specifically, for all probes on the array, we look up all other probes within 100K nucleotides upstream and downstream, and average the p-values thus obtained. The PValue, CloseProbes and MeanPValue columns contain the p-value for the "center" probe, the number of probes within the 100K window and the average p-values for all those probes, respectively. The averaged p-values have no statistical meaning, but can be used as an indicator for regions of interest, which we call "methylation hotspots". We recommend opening this file in a spreadsheet program and setting appropriate sort orders and filters, such as a descending average p-value sort and a filter to keep only regions where a significant number of probes are present (such as >5).

5.4 Enrichment analysis of probe categories

The analysis pipeline performs simple tests to determine if given categories of probes show enrichment in the set of all identified DMRs compared to the set of all probes on the EDMA microarray. This produces a text file (Enrich.txt) as well as the set of plots presented below. For each given categorization of probes (CpG island density, CpG island length, CpG island neighborhood, genic region, repeated elements present), three graphs are produced:

- The first graph (Figure 10) presents the absolute proportion of probes in each categories within (i) the set of all probes and (ii) the set of all differentially methylated probes.
- The second graph (Figure 11) is another way of representing the same data as the first graph. It shows the $\log_2$ of relative enrichment of each category within the set of differentially methylated regions. For example, a bar of length 1 would mean that there are two times as many probes of that category in the set of all DMRs as would be expected if DMRs had been selected randomly amongst all probes on the array; a bar length of -1 would mean that there are only half as many as would be expected. The farther left a bar extends, the more methylation tends to be conserved within this category. The farther right, the more changes in methylation are concentrated in this category. A bar near zero does not mean that the category shows no changes in methylation, but only that the number of changes is in range with what would be expected by chance. The numbers besides a bar represents how many DMRs fit in that category, and the percentage of all DMRs that this represents.
- The third figure shows to what extent the various categories are split between hypermethylated in the reference condition and hypermethylated in the other condition (or, equivalently, hypermethylated and hypomethylated in the reference condition). The first bar represents the proportion for all probes, and the dotted line represents this baseline value.

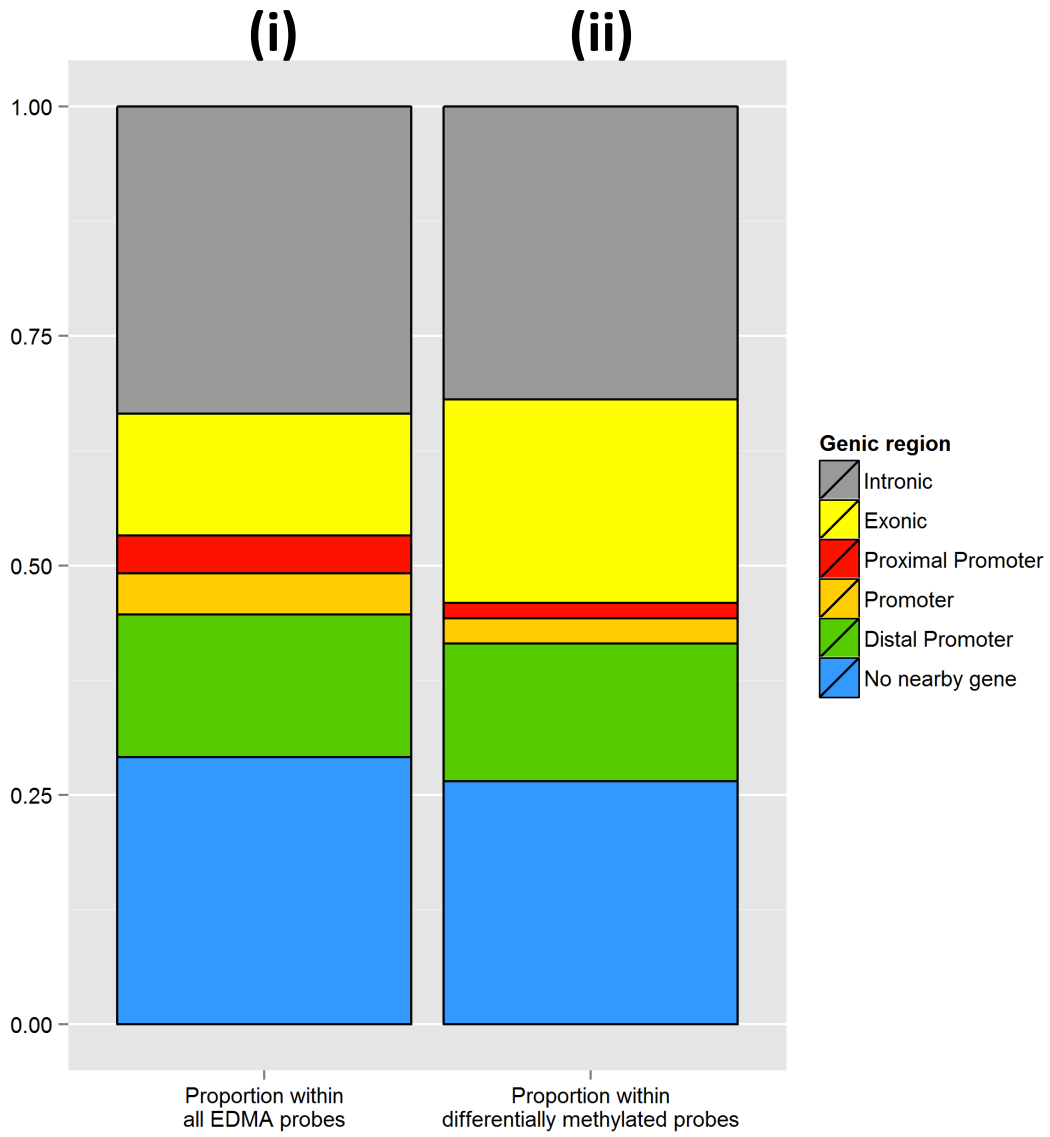

Figure 10. Stacked plot of category enrichment. The graph represents the proportion of probes in each categories within (i) the set of all probes and (ii) the set of all differentially methylated probes.

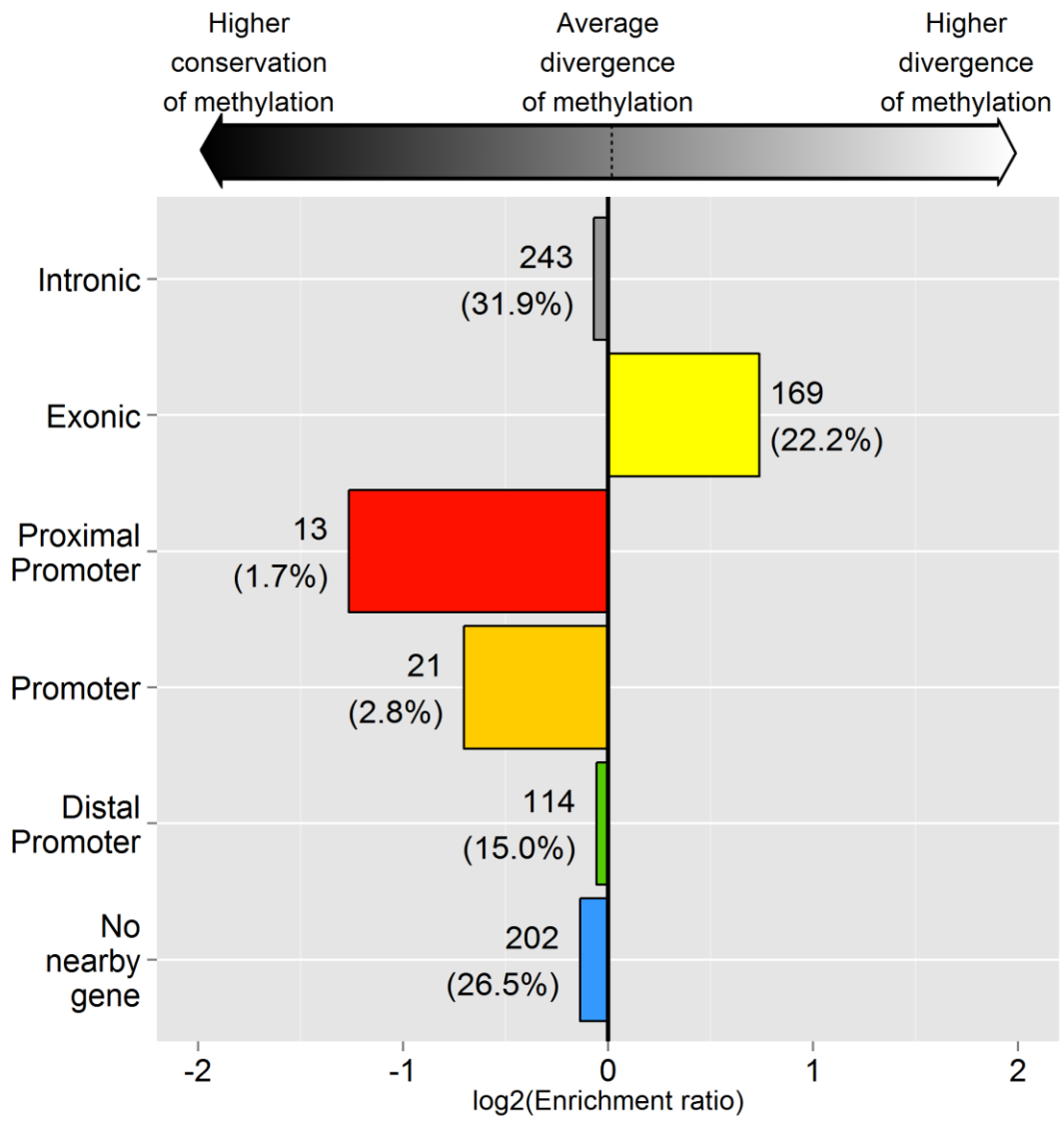

Figure 11. Bar plot of the log2 of relative category enrichments. A bar of length 1 would mean there are two times as many differentially expressed probes of that category than what would be expected if differentially methylated probes had been selected randomly; a bar length of -1, would mean there are only half as many as would be expected. The numbers besides a bar represents how many DMRs fit in that category, and the percentage of all DMRs that this represents.

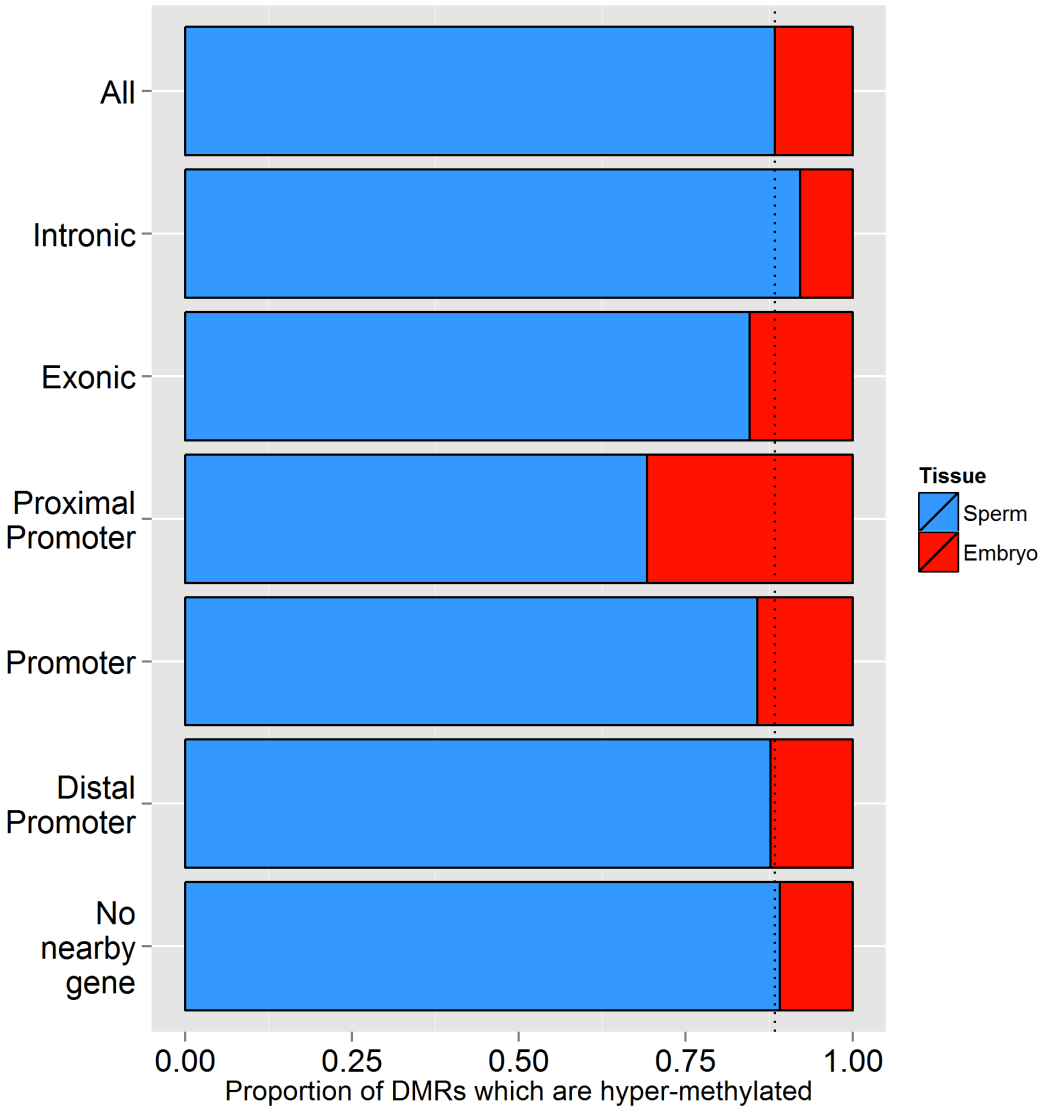

Figure 12. Proportions of DMRs which are found to be hypermethylated in each tissue, split by category. The dotted line represents the baseline of this ratio when all probes are taken into account, regardless of category.

5.5 Circular plot

For each experiment, the analysis pipeline produces three circular plots:

- X.legend.png, which summarizes and presents all EDMA probes.
- X-Significant.png, which summarizes all identified DMRs.
- X-30K.legend.png, which sets a arbitrary fold-change and p-value thresholds at the 30,000th most significant elements of the EDMA array. That is to say, all absolute fold-changes and p-values are sorted in ascending and descending orders, respectively, and the 30,000th element of both lists are identified. These values are then used as cutoff for filtering all probes. This is done to generate probe lists that are ~15,000 elements long and represent the most variable elements on the array, for when the list of DMRs is too sparse to produce a nice genome-wide plot.

Furthermore, EMAP can produce two types of circular plots, according to the types of data which are provided. If both transcriptomic and epigenetic data are provided, EMAP produces a combined plot (section 5.5.1). If only epigenetic data is available, EMAP produces a standalone epigenetic plot (section 5.5.2). All layers of all plots show values within windows of 5,000 bases. P-value layers have an overlay representing the location of the 100 most differentially methylated/expressed probes. Hyper-methylated/Over-expressed probes are shown as upward yellow arrows; hypo-methylated/Under-expressed probes are shown as downward red arrows.

5.5.1 Combined circular plot

Combined analyses (Transcriptomic + epigenetic, figure 13) produce a circular plot presenting 5-layers of data. Do note that the filtering steps for the -Significant and -30K plots only apply for the epigenetic results.

1. Epigenetic p-values
2. Epigenetic fold-changes
3. Transcriptomic p-values
4. Transcriptomic fold-changes
5. A list of positioned genes whose transcriptomic and epigenetic changes vary in opposite directions, referred to as "Concordant changes".

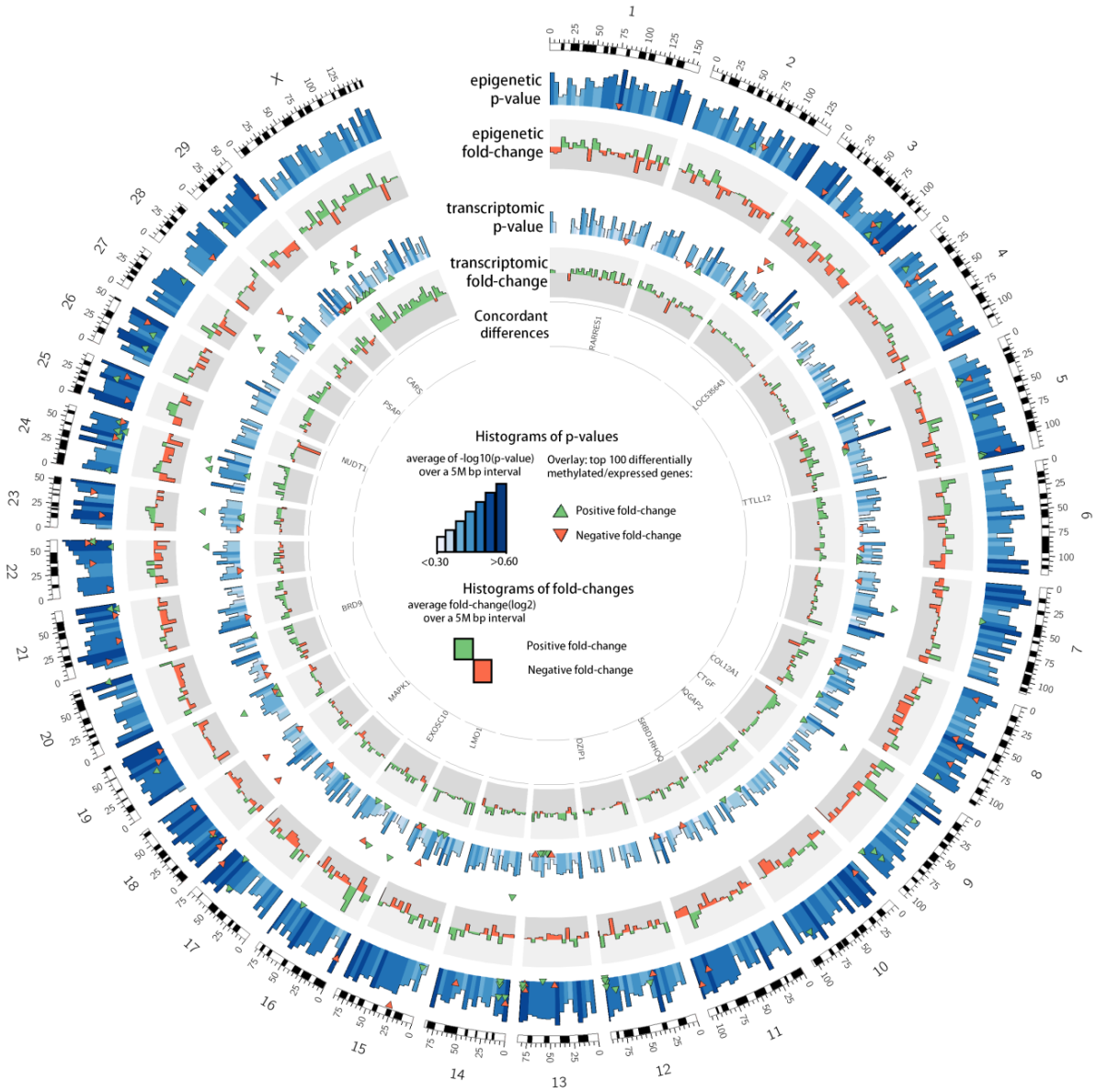

Figure 13. Combined circular plot.

5.5.2 Standalone epigenetic circular plot

For epigenetic experiments without a transcriptomic counterpart, the circular plot presents the following layers:

1. Epigenetic p-values
2. Epigenetic fold-changes
3. Mean intensity for the reference condition
4. Fold-changes of imprinted genes
5. Gene-symbols of imprinted genes presented on layer 4.

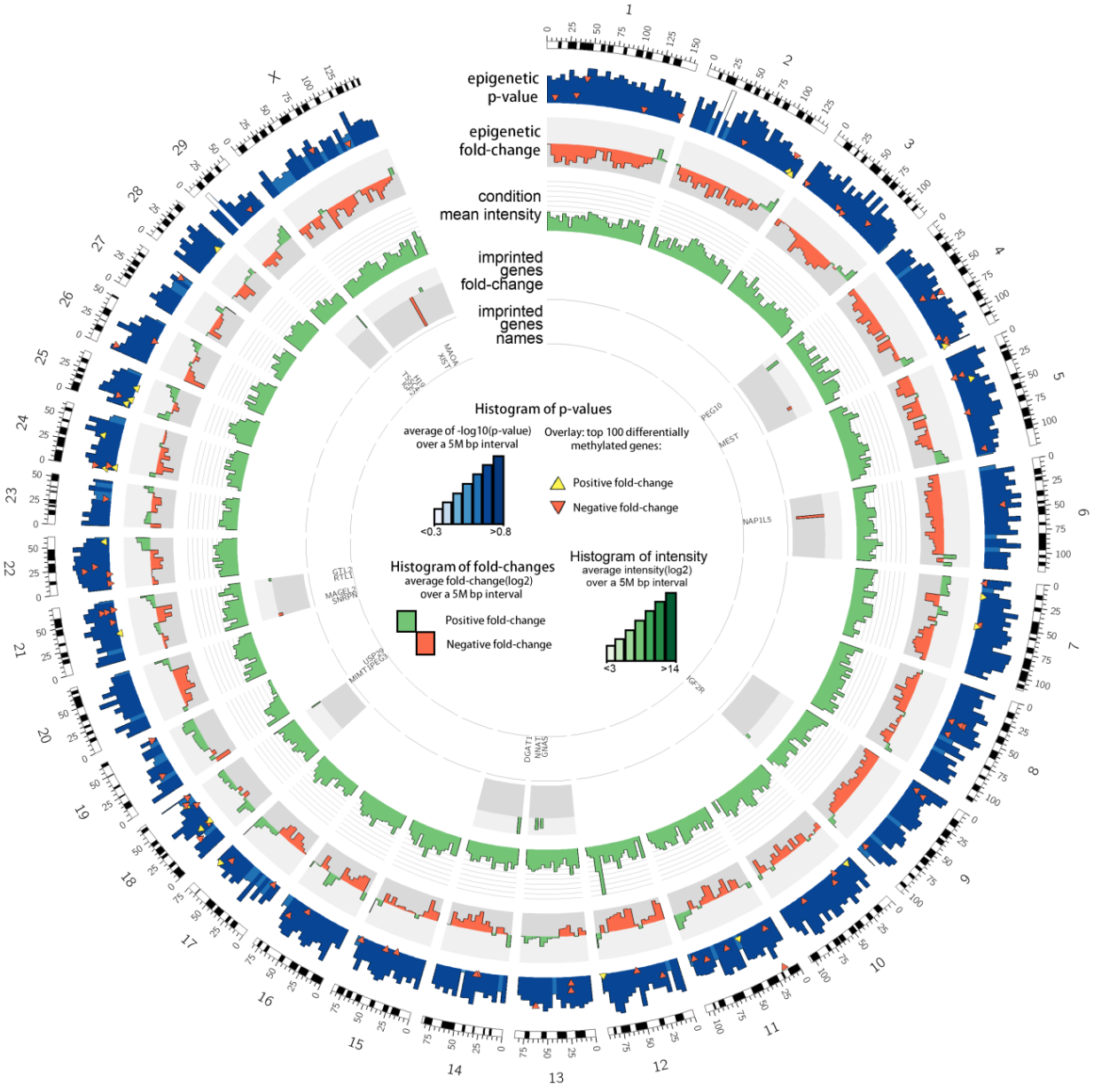

Figure 14. Circular plot for epigenetic analysis without an accompanying transcriptomic analysis.

6. Frequently Asked Questions (FAQ)

Q: How can I know if a genomic region is methylated?

A: You can compare the signal for that region with the detection cutoff (See section 4.1). While being above the cutoff is not a **guaranteed** sign of methylation, being consistently above it in a given condition is a strong indicator of possible methylation. Note however that being below the cutoff is **not** indicative of a complete lack of methylation. An MseI-MseI fragment might contain numerous sites targeted by the methyl-sensitive enzymes, and all of those sites must evade digestion for the fragment to be exponentially amplified. Thus a fragment with five methyl-sensitive sites, four of which are methylated, could still end up with a signal below the cutoff.

Q: My exogenous spike controls are in my list of DMRs. What does this mean?

This might mean that either your starting quantities of DNA were uneven, or that one of your tissues shows such hypomethylation compared to the other that the sensitive digestion caused an uneven bias in the first or second round of PCR.

Q: How can I verify if my favorite imprinted gene shows change in methylation?

Determine the genomic coordinates which act as imprinting controls, and find the probe(s) surveying these coordinate. Be aware that since the EDMA platform look for methylation changes in broad regions, fine-grained methylation control (based on changes in only one or two CpG dinucleotides) might escape detection.

Q: What should I do with my list of DMRs?

EMAP provides a wide-range of pre-packaged analysis options, including hotspot detection, a file for importing into IPA, files for visualizing results in a genome browser, category enrichment, and a list of genes showing concordant canonical changes in both the epigenome and transcriptome. Beyond that, you will have to survey your region list the old fashioned way and use your knowledge of the underlying biology to make sense of the data.

Q: What are the default fold-change and p-value cutoffs for identifying DMRs?

EMAP uses an absolute fold-change of log2(1.5) and a p-value of 0.05 as defaults cutoffs for DMR identification.