INTRODUCTION

DNA methylation in vertebrates is characterized by the addition of a methyl or hydroxymethylgroup to the C5 position of cytosine, which occurs mainly in the context of CG dinucleotides.

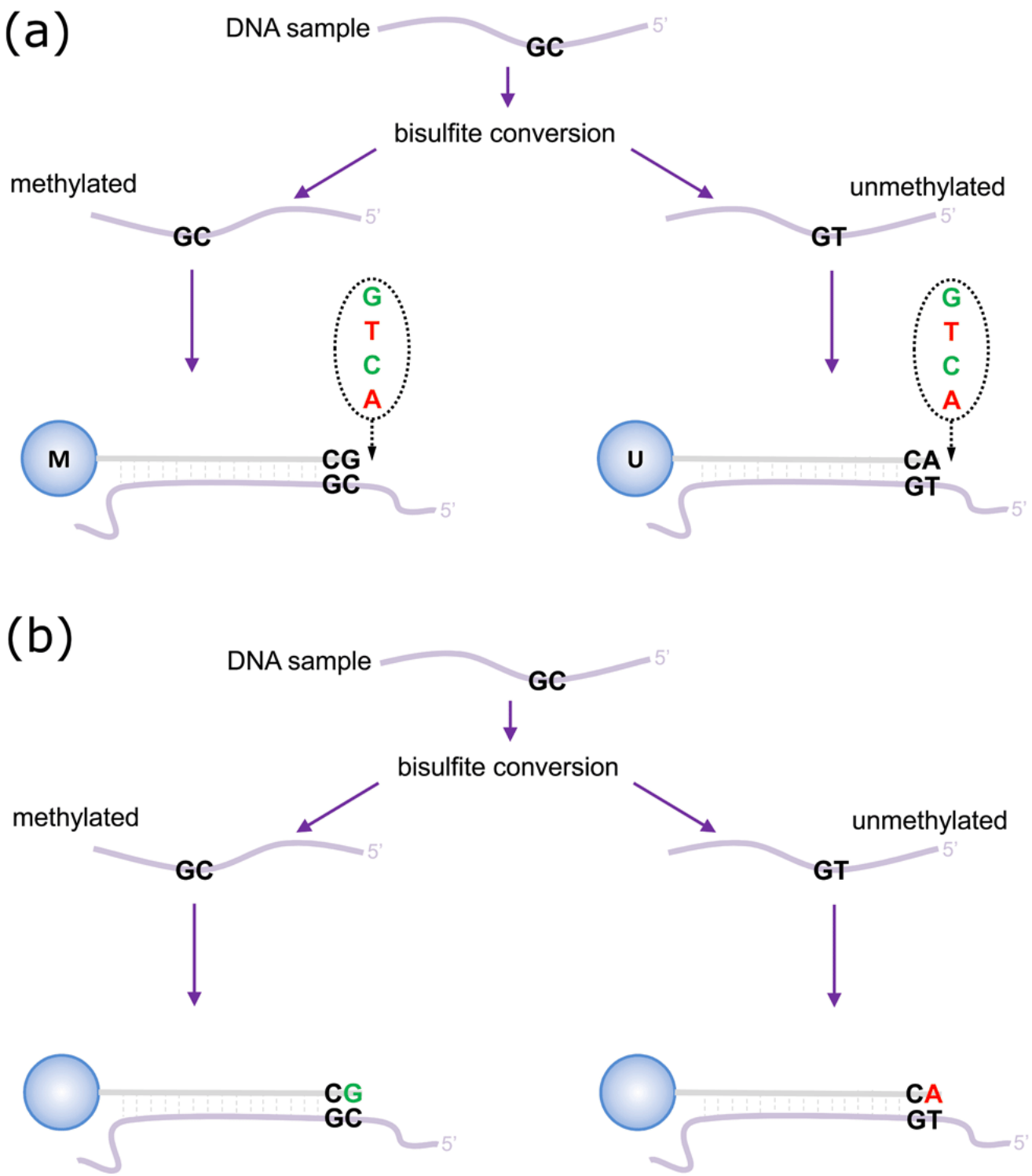
The bisulfite treatment of DNA mediates the deamination of cytosine into uracil, and these converted residues will be read as thymine, as determined by PCR-amplification and subsequent Sanger sequencing analysis.

However, 5 mC residues are resistant to this conversion and, so, will remain read as cytosine. Thus, comparing the Sanger sequencing read from an untreated DNA sample to the same sample following bisulfite treatment enables the detection of the methylated cytosines. Like bisulfite sequencing, the Illumina Infinium assay detects methylation status at single base resolution.

Regardless of the Illumina array version, for each CpG, there are two measurements: a methylated intensity (denoted by M) and an unmethylated intensity (denoted by U).

These intensity values can be used to determine the proportion of methylation at each CpG locus. Methylation levels are commonly reported as either beta values ($\beta = M/(M + U)$) or M-values ($M \text{ value} = \log2(M/U)$).

Beta values are generally preferable for describing the level of methylation at a locus or for graphical presentation because percentage methylation is easily interpretable. However, due to their distributional properties, M-values are more appropriate for statistical testing.



(**a**) Infinium I assay. Each individual CpG is interrogated using two bead types: methylated (M) and unmethylated (U). Both bead types will incorporate the same labeled nucleotide for the same target CpG, thereby producing the same color fluorescence. The nucleotide that is added is determined by the base downstream of the "C" of the target CpG. The proportion of methylation can be calculated by comparing the intensities from the two different probes in the same color.
(**b**) Infinium II assay. Each target CpG is interrogated using a single bead type. Methylation state is detected by single base extension at the position of the "C" of the target CpG, which always results in the addition of a labeled "G" or "A" nucleotide, complementary to either the "methylated" C or "unmethylated" T, respectively. Each locus is detected in two colors, and methylation status is determined by comparing the two colors from the one position.

Each CpG is associated with two measurements: a methylated measurement and an “un”-methylated measurement. These two values can be measured in one of two ways: using a “Type I” design or a “Type II design”.

The Green intensity measures the methylated signal, and the Red intensity measures the unmethylated signal.

CpGs measured using a **Type I** design are measured using a single color, with two different probes in the same color channel providing the methylated and the unmethylated measurements.

CpGs measured using a **Type II** design are measured using a single probe, and two different colors provide the methylated and the unmethylated measurements. Practically, this implies that on this array there is not a one-to-one correspondence between probes and CpG positions. The EPIC array has 8 arrays per slide and 64 arrays per plate.

The EPIC arrays use both TypeI and TypeII designs

QUALITY CONTROL

We can generate a detection p-value for every CpG in every sample, which is indicative of the quality of the signal.

The method used by minfi to calculate detection p-values compares the total signal (M + U) for each probe to the background signal level, which is estimated from the negative control probes.

Very small p-values are indicative of a reliable signal whilst large p-values, for example >0.01, generally indicate a poor quality signal.

NORMALIZATION

Although there is no single normalisation method that is universally considered best, a recent study has suggested that a good rule of thumb within the minfi framework is that the `preprocessFunnorm()` function is most appropriate for datasets with global methylation differences such as cancer/normal or vastly different tissue types, whilst the `preprocessQuantile()` function is more suited for datasets where you do not expect global differences between your samples, for example a single tissue.

MDS Plots (PCA)

MDS plots are based on principal components analysis and are an unsupervised method for looking at the similarities and differences between the various samples.

FILTERING

Poor performing probes are generally filtered out prior to differential methylation analysis. We filter out probes that have failed in one or more samples based on detection p-value.

We can also perform the removal of probes where common SNPs may affect the CpG.

We will also filter out probes that have shown to be cross-reactive, that is, probes that have been demonstrated to map to multiple places in the genome.

PROBE-WISE DIFFERENTIAL METHYLATION ANALYSIS

A convenient way to set up the model when the user has many comparisons of interest that they would like to test is to use a contrasts matrix in conjunction with the design matrix.
A contrasts matrix will take linear combinations of the columns of the design matrix corresponding to the comparisons of interest.

Since we are performing hundreds of thousands of hypothesis tests, we need to adjust the p-values for multiple testing.

A common procedure for assessing how statistically significant a change in mean levels is between two groups when a very large number of tests is being performed is to assign a cut-off on the false discovery rate. Typically 5% FDR is used, and this is interpreted as the researcher willing to accept that from the list of significant differentially methylated CpG sites, 5% will be false discoveries.

DIFFERENTIAL METHYLATION ANALYSIS OF REGIONS

We will perform an analysis using the `dmrcate`. As it is based on `limma`, we can directly use the design and contMatrix we previously defined.

The function `cpg.annotate()` annotates CpGs with their chromosome position and statistical test.

GENE ONTOLOGY TESTING

There may be many thousands of CpGs significantly differentially methylated. In order to gain an understanding of the biological processes that the differentially methylated CpGs may be involved in, we can perform gene ontology or KEGG pathway analysis using the `gometh()` function in the `missMethyl` package

It is important to keep in mind that we are not observing gene level activity such as in RNA-Seq experiments, and that we have had to take an extra step to associate CpGs with genes.

DIFFERENTIAL VARIABILITY

Rather than testing for differences in mean methylation, we may be interested in testing for differences between group variances.

Including Plots

You can also embed plots, for example:

```
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.