MassArray Analytical Tools

Reid F. Thompson and John M. Greally

April 22, 2010

Contents

1 Introduction

The *MassArray* package provides a number of tools for the analysis of MassArray data, with particular application to DNA methylation and SNP detection. The package includes plotting functions for individual and combined assays, putative fragmentation profiles, and potential SNP locations level data useful for quality control, as well as flexible functions that allow the user to convert probe level data to methylation measures.

In order to use these tools, you must first load the *MassArray* package:

```
> library(MassArray)
```

It is assumed that the reader is already familiar with mass spectrometry analysis of base-specific cleavage reactions using the Sequenom workflow and MassCLEAVE assay. If this is not the case, consult the Sequenom User's Guide for further information (Sequenom, 2008).

Throughout this vignette, we will be exploring actual data for 2 amplicons, each containing a set of samples.

2 Changes for MassArray in current BioC release

- This is the first public release of the *MassArray* package.

3 Optimal Amplicon Design

The `ampliconPrediction()` function is designed to take an input nucleotide sequence and output a graphic depiction of the pattern of individual CGs that are measurable by any of the four possible RNase reaction/strand combinations (either a T or C reaction for either the forward or reverse strands). The function also returns a tabular output of the measurability of each CG in the amplicon under each of the four possible conditions.

This functionality will be demonstrated for a short nucleotide sequence containing two different CG sites. This sequence is far too short for use as a successful MassArray assay; however, this example should be illustrative of the differences between reactions. Both the graphical and tabular depictions are shown here, although similar representations throughout the remainder of this documentation will only show graphical output.

```
> results <- ampliconPrediction("AAAATTTTCCCCTCTGCGTGAGAGAGTTGTCCGACAAAA")
> results

$summary
  required summary   T+    C+    T-    C-
1    FALSE    TRUE TRUE FALSE FALSE  TRUE
2    FALSE    TRUE TRUE FALSE  TRUE FALSE

$counts
         summary + T+ C+ - T- C-
all            2 2  2  0 2  1  1
required       0 0  0  0 0  0  0
```

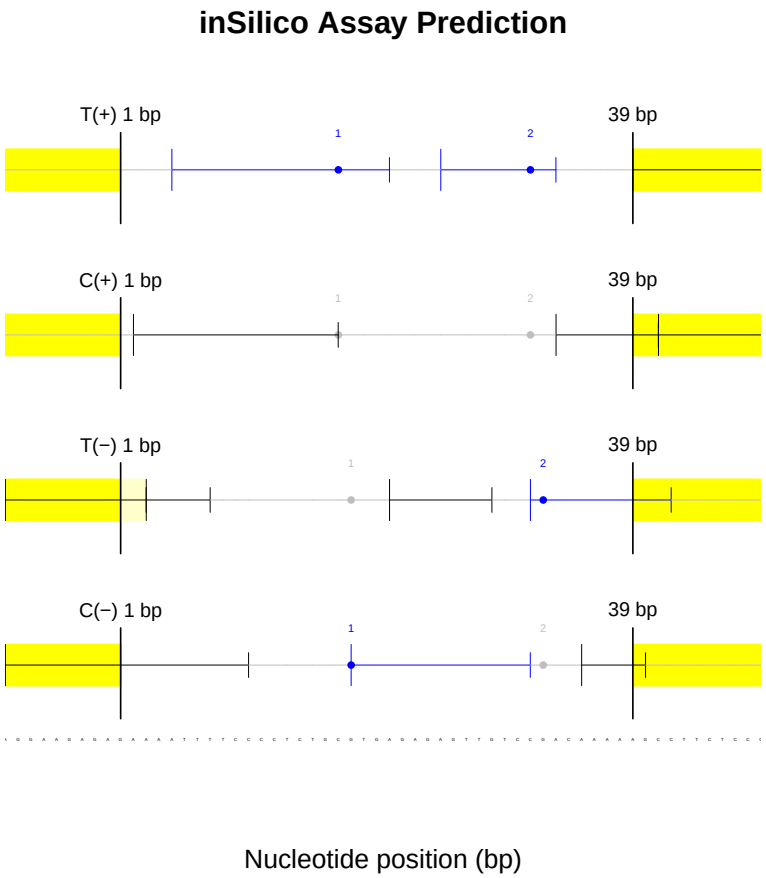

Figure 1: Basic amplicon prediction for a short sequence containing two CG sites. Putative fragmentation patterns are shown for T and C-cleavage reactions on both the plus and minus strands. CG dinucleotides (filled circles) are numbered and color-coded according to their ability to be assayed: fragment molecular weight outside the testable mass window (gray), uniquely assayable site (blue). Yellow highlights represent tagged sequences.

You may also highlight specific sequences (or individual CG dinucleotides) using paired parentheses or arrowheads in the input nucleotide sequence. In the following example, the second CG dinucleotide (CG#2) is marked/highlighted in lavendar as an important site of analysis. A four basepair "primer" is labeled on each end of the input sequence using arrowheads.

```
> ampliconPrediction("AAAA>TTTTCCCCTCTGCGTGAGAGAGTTGTC(CG)AC<AAAA")
```

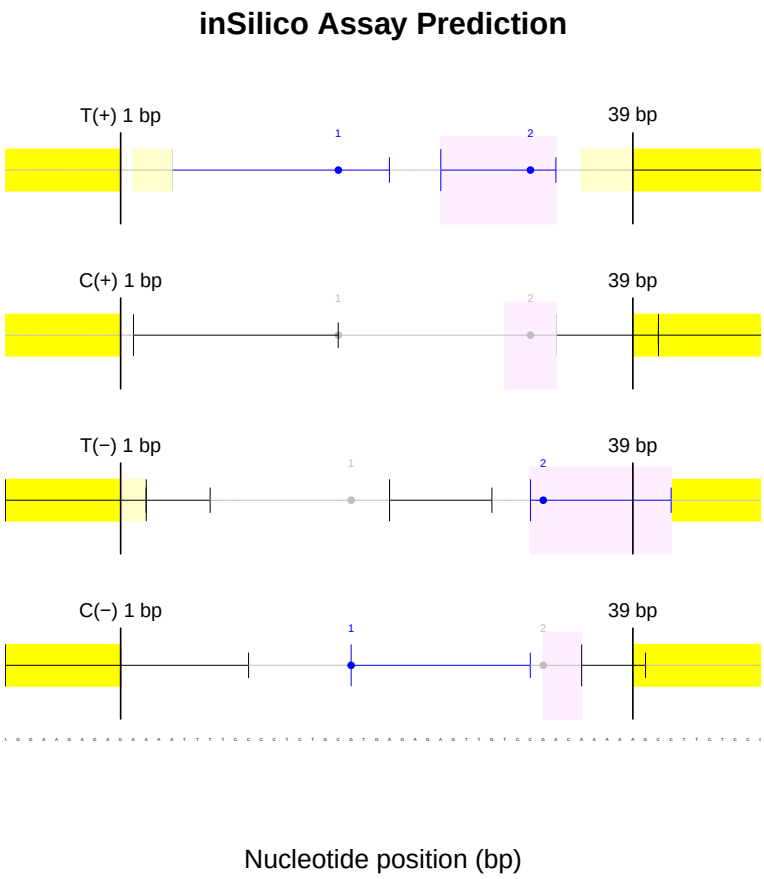

Figure 2: Basic amplicon prediction for a short sequence containing two CG sites, the second of which is highlighted. Putative fragmentation patterns are shown for T and C-cleavage reactions on both the plus and minus strands for the specified sequence. CG dinucleotides (filled circles) are numbered and color-coded according to their ability to be assayed, where gray indicates that the CG is located on a fragment whose molecular weight is outside the usable mass window and blue indicates a uniquely assayable site. Yellow highlights represent tagged/primer sequences, while lavender highlights denote user-specified "required" sites.

However, the large majority of assays are complicated by molecular weight overlaps between CG-containing fragments and one or more non CG-containing fragments, or alteratively between more than one CG-containing fragment. These overlaps reduce the ability to resolve independent methylation status for a given site, and thus are flagged in red to alert the user to this situation. All CG sites or fragments marked in red contain some form of molecular weight overlap with one or more other fragments. CG sites that are linked by gray arrows indicate fragments that have molecular weight overlap(s) with other CG-containing fragments. Methylation status at these labeled sites cannot be measured independently from the methylation status of overlapping CG dinucleotides using the indicated reaction/strand combination(s). A simple example of such an assay is shown below.

```
> ampliconPrediction("AAAATTTTCCCCTCTGCGTGAGAGAGTTGTC(CG)ACTTCCCCTCTGCGTGAGAGAGTTGTCCGACAAAA")
```

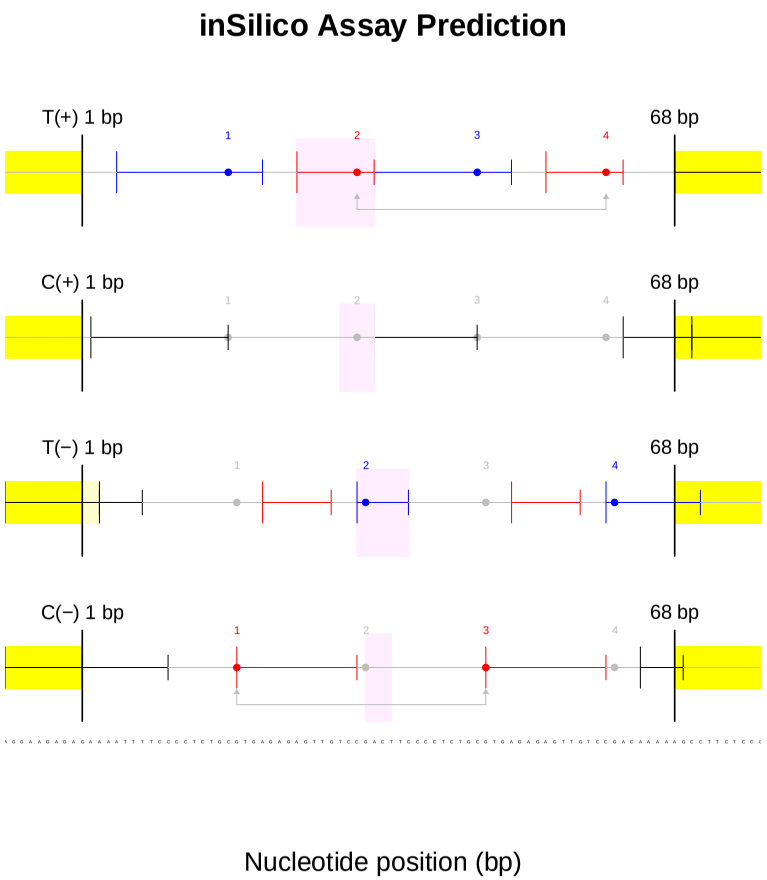

Figure 3: Basic amplicon prediction for a short sequence containing four CG sites with some molecular weight overlaps. Putative fragmentation patterns are shown for T and C-cleavage reactions on both the plus and minus strands for the specified sequence. CG dinucleotides (filled circles) are numbered and color-coded according to their ability to be assayed, where gray indicates that the CG is located on a fragment whose molecular weight is outside the usable mass window, red indicates a molecular weight overlap with another fragment, and blue indicates a uniquely assayable site. Linked arrowheads denote molecular weight overlaps between multiple CG-containing fragments. Yellow highlights represent tagged/primer sequences, while lavender highlights denote user-specified ”required” sites.

4 Conversion Controls

Accurate measurement of methylation status presupposes that the bisulphite conversion reaction runs to completion. If, however, bisulphite conversion is incomplete, "methylation" measured at any CG dinucleotide will be composed of actual methylation mixed with signal from remnant unconverted, unmethylated cytosines. For many target regions, this issue is mitigated by selective amplification of fully-converted templates (primers should contain at least 4 non-CG 'C's). Nevertheless, amplicons may still contain some background level of partially unconverted DNA: primer selection criteria occasionally need to be relaxed, and PCR tends to enrich underrepresented sequences.

In order to measure levels of unconverted non-CG cytosines in a given MassArray sample, we implemented a conversion control measurement function (`evaluateConversion()`, which is itself a wrapper function for `convControl()`), to search the predicted fragmentation profile for non-CG cytosines that occur in the absence of CG dinucleotides. Moreover, potential conversion controls are automatically filtered to remove any molecular weight overlaps with other predicted fragments, so that they may be considered in isolation.

The identification of fragments usable as conversion controls occurs during amplicon prediction for a given sequence input. Approximately 91% of assays are likely to contain such conversion controls, however, some amplicons may lack this particular measureable (Thompson et al., 2009). Called as an accessory function, `convControl()` defines conversion control fragments wherever they meet the following criteria:

1. sequence containing no CGs
2. sequence containing at least one non-CG cytosine
3. sequence containing at least one TG
4. molecular weight within the useable mass window
5. no molecular weight overlap with other predicted fragments
6. no molecular weight overlap of sequence containing one unconverted cytosine with other predicted fragments

Here is an example of usable conversion controls identified in a given amplicon. Three reactions contain measureable conversion controls (labeled in green), however, one reaction lacks usable conversion controls as shown.

```
> ampliconPrediction("CCTGTCCAGGGGCACTCCATATTTTCCTACCTGTCCCTCTTTGCTTGTAAAAACAAATTAAACAGGGATCCCAGCAACTTCG")
```

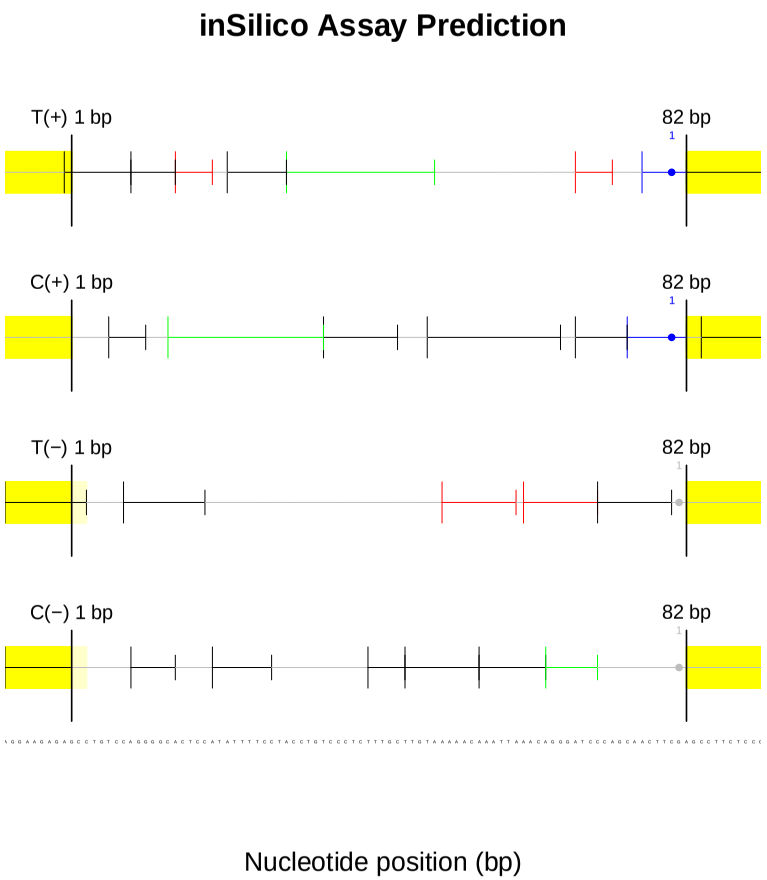

Figure 4: Basic amplicon prediction for a short sequence showing the identification of usable conversion controls. Putative fragmentation patterns are shown for T and C-cleavage reactions on both the plus and minus strands for the specified sequence. CG dinucleotides (filled circles) are numbered and color-coded according to their ability to be assayed, where gray indicates that the CG is located on a fragment whose molecular weight is outside the usable mass window, red indicates a molecular weight overlap with another fragment, and blue indicates a uniquely assayable site. Fragmentation patterns are shown in corresponding colors, with the addition of green fragments indicating usable conversion controls. Yellow highlights represent tagged/primer sequences.

Those fragments meeting the above criteria are flagged appropriately and are treated as if they contained a CG, thus enabling the software to determine the extent of the bisulphite conversion reaction. We then applied this tool to data from a number of samples and show detected levels of unconverted cytosines for two examples. For the large majority of samples, we find that bisulphite conversion is near-complete (ranging from 98-100%). However, we also show significant retention of unconverted cytosines for an amplicon (rat chr17:48916975-48917295, rn4 Nov. 2004 assembly, UCSC Genome Browser) that approaches 25%. This incomplete conversion is a relatively rare experimental outcome in our experience, nevertheless it demonstrates the necessity of a conversion control measure as part of each experiment to ensure accuracy of the data (Thompson et al., 2009).

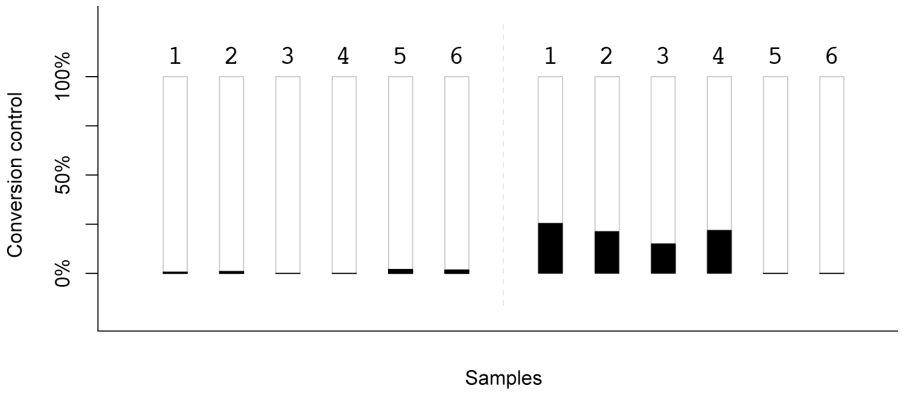

Figure 5: Measurement of conversion controls for two amplicons shows variable extent of bisulphite conversion. Conversion controls were measured for two different amplicons using two sets of six biological replicates (each, the product of a separate bisulphite conversion reaction); samples (labeled 1-6) are shown here divided by amplicon (one set at left, one set at right). Bar height (black) depicts the percentage of unconverted cytosines detected, with 0% indicating complete conversion of measured cytosines and 100% indicating a complete failure of conversion. Four of six samples among the second set (at right) show significant retention of unconverted cytosines (numbered 1-4), as measured by conversion controls.

5 Data Import

In order for MassArray data to be loaded into an R workspace using this software, it must first be exported from Sequenom's EpiTyper in a tab-delimited file format. Please note that it is imperative that a single amplicon be exported at a single time, otherwise the import will fail. Once you have successfully opened an individual amplicon in the EpiTyper application, enable display of the 'Matched Peaks' tab, both the 'T' and 'C' reactions, and the 'Show All Matched Peaks' and 'Show All Missing Peaks' options. Data must be exported using the 'Export Grid' command in a tab-delimited text format. For further clarification and support, consult the EpiTyper Application Guide provided with the software or available from the Sequenom website.

Once data has been properly exported, it may be loaded into an open R workspace. An example of the import commands for an amplicon with two samples (A and B) is shown here:

```
> sequence <- "CCAGGTCCAAAGGTTCAGACCAGTCTGAA>CCTGTCCAGGGGCACTCCATATTTTCC"
> sequence <- paste(sequence, "TACCTGTCCCTCTTTGCTTGTAAAAACAAATTAAACAGGGA",
+     sep = "")
> sequence <- paste(sequence, "TCCCAGCAACTTCGGGGCATGTGTGTAACTGTGCAAGGAGC",
+     sep = "")
> sequence <- paste(sequence, "GCGAAGCCCAGAGCATCGCCCTAGAGTTCGGGCCGCAGCTG",
+     sep = "")
> sequence <- paste(sequence, "CAGAGGCACATCTGGAAAAGGGGGAGGGGTCGAAGCGGAGG",
+     sep = "")
> sequence <- paste(sequence, "GGACAAGAAGCCCCCAAACGACTAGCTTCTGGGTGCAGAGT",
+     sep = "")
> sequence <- paste(sequence, "CTGTGTCAC(CG)GGGGTTAGTTACCTGTCCTACGTTGATG",
+     sep = "")
> sequence <- paste(sequence, "AATCCGTACTTGCTGGCTATGCGGTCTGCCTCCGCGAATCC",
+     sep = "")
> sequence <- paste(sequence, "GC(CG)GC<GATCTTCACTGCCCAGTGGTTGGTGTA",
+     sep = "")
> data <- new("MassArrayData", sequence, file = "Example.txt")

Performing inSilico Fragmentation: T, C ... FINISHED
Importing matched peaks file (Example.txt):
        Reading assay (Example), 2T+0C rxns (EpiTyper v1.0.5):
                T reaction:
                        Reading sample (A) ... FINISHED
                        Reading sample (B) ... FINISHED
        Analyzing conversion control(s) ... FINISHED
        Estimating primer dimer level(s) ... FINISHED
        Estimating adduct level(s) ... FINISHED
        Analyzing CpG methylation ... FINISHED
```

Note that additional options may be specified. For more information, consult the description of the `MassArrayData-class` in the R help documentation. Also, note that the `MassArrayData-class` is a structural composite of many individual pieces of data describing everything from the fragmentation structure (`MassArrayFragment-class`) to the actual MassArray peak information (`MassArrayPeak-class`, which taken together form spectral data – see `MassArraySpectrum-class`). The user need not concern themselves with these particulars as data import creates the relevant structures automatically. However, individual data elements may be interrogated or modified through knowledge of this structure which may prove useful in certain circumstances after data has already been imported. Please consult the relevant help files for a further discussion of each of these detailed data structures.

6 Data Visualization

We have implemented a visual alternative to the epigram that takes methylation data and displays it in the form of color-filled bars, where the shaded height indicates percent methylation.

```
> plot(data, collapse = FALSE, bars = FALSE, scale = FALSE)
```

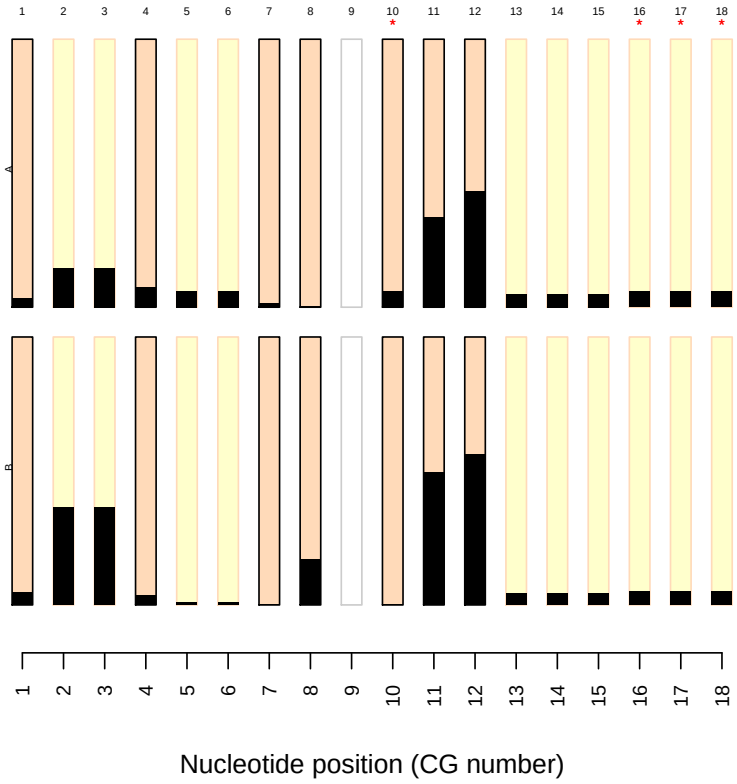

Figure 6: Basic plotting tool for MassArray data (individual samples). Bar height denotes percent methylation on a scale from 0% (low) to 100% (high) for each CG (eighteen of which are shown here in order from left to right). Note red stars indicate user-defined "required" sites. CG dinucleotides located on a fragment with other CGs are indicated as bars with yellow background. CG sites that are putatively outside the usable mass window are shown in gray outline.

This graphical depiction also includes the ability to display "error bars" which correspond to the median absolute deviation as a measure of variability among a collection of measurements. The methylation data displayed represents the average among samples.

```
> plot(data, collapse = TRUE, bars = TRUE, scale = FALSE)
```

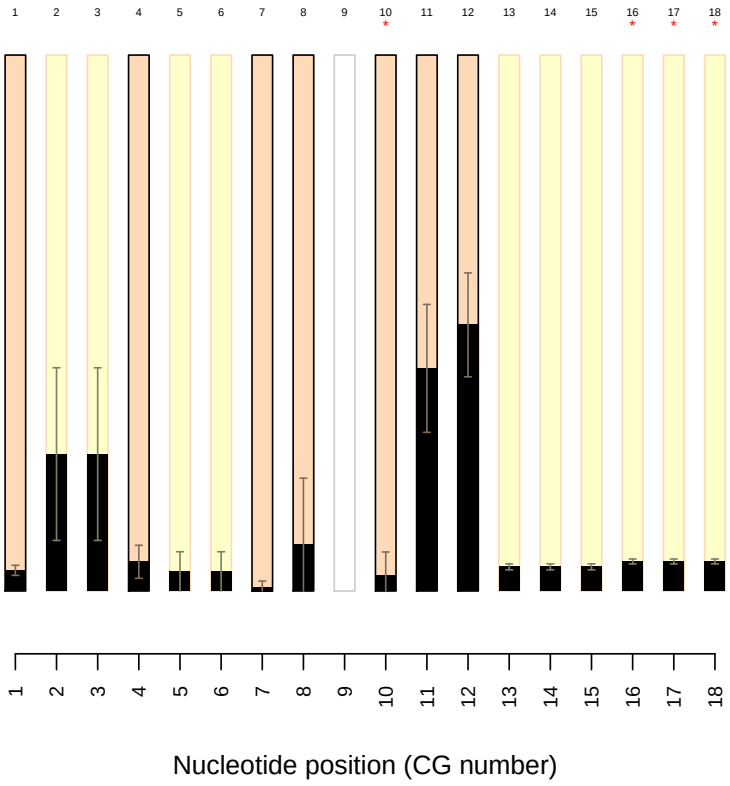

Figure 7: Basic plotting tool for MassArray data (data averaged across samples). Bar height denotes percent methylation on a scale from 0% (low) to 100% (high) for each CG (eighteen of which are shown here in order from left to right), with error bars indicating median absolute deviation. Note red stars indicate user-defined "required" sites. CG dinucleotides located on a fragment with other CGs are indicated as bars with yellow background. CG sites that are putatively outside the usable mass window are shown in gray outline.

The data can also be displayed in a positionally-informative manner, wherein the x axis represents relative nucleotide position for each measured CG. Display of error bars may be optionally deactivated.

```
> plot(data, collapse = TRUE, bars = FALSE, scale = TRUE)
```

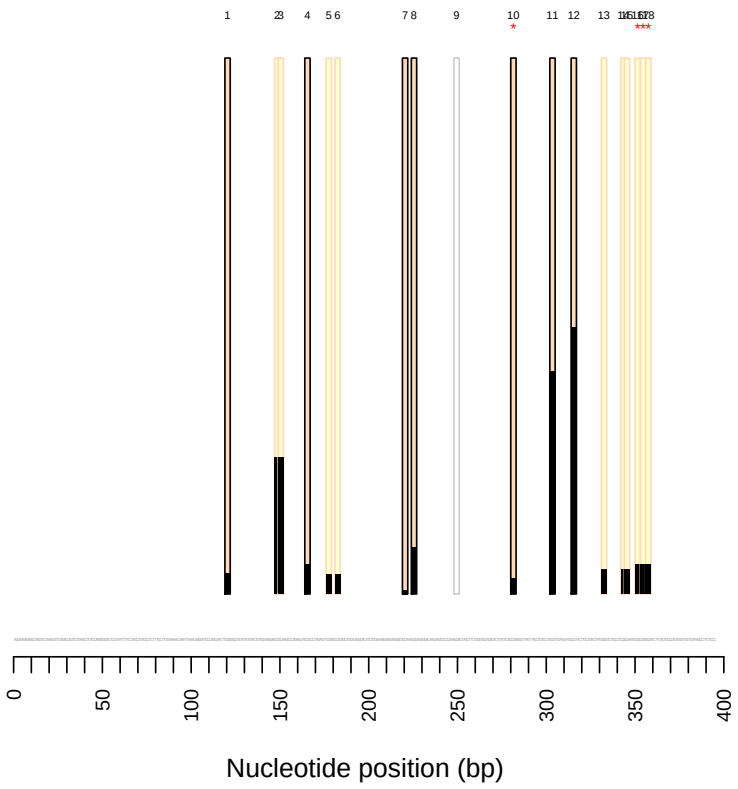

Figure 8: Basic plotting tool for MassArray data (data scaled to relative nucleotide positions). Bar height denotes percent methylation on a scale from 0% (low) to 100% (high) for each CG (eighteen of which are shown here in order from left to right), with error bars indicating median absolute deviation. Note red stars indicate user-defined "required" sites. CG dinucleotides located on a fragment with other CGs are indicated as bars with yellow background. CG sites that are putatively outside the usable mass window are shown in gray outline.

7 Single Nucleotide Polymorphism Detection

Single nucleotide polymorphisms (SNP) can interrupt the ability to interpret methylation status at one or more CG dinucleotides. This R package thus enables the identification of putative SNPs by comparison of expected and observed data. Any mismatches may be analyzed using an exhaustive string substitution approach (`identifySNPs()`), where each existent base pair in the sequence is substituted with one of the three remaining bases or a gap (representing a single base-pair deletion) and then judged for its ability to modify the expected fragmentation pattern in a manner that matches the observed data. The `identifySNPs()` function, however, serves as an internal method. The best way to access the built-in SNP dectection is via the `evaluateSNPs()` function.

```
> SNP.results <- evaluateSNPs(data)
```

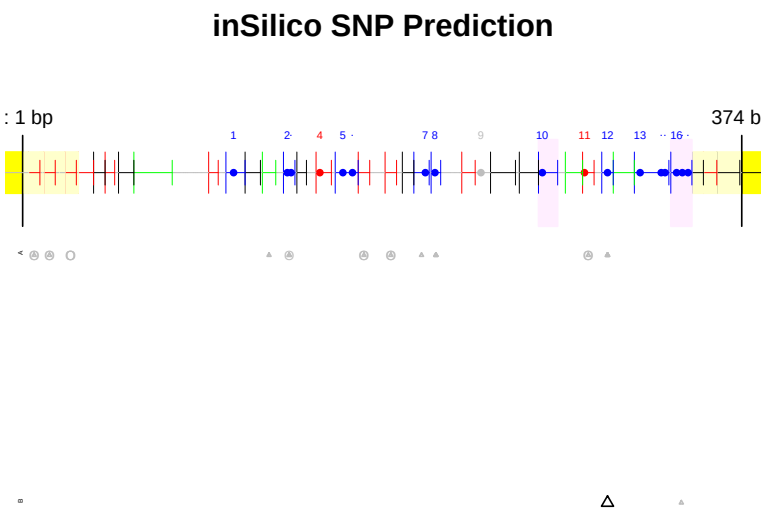

Figure 9: Graphical output from putative SNP detection. The T-cleavage fragmentation profile is shown (top panel). CG dinucleotides (filled circles) are numbered and colored in blue. Other fragments are colored according to their ability to be assayed: fragment molecular weight outside the testable mass window (gray), fragment molecular weight overlapping with another fragment (red), fragment containing a potential conversion control (green), or fragment uniquely assayable but containing no CGs (black). Putative SNPs are shown directly below their location within the amplicon fragmentation profile. Each row represents analysis from a single sample (two different biological samples). Small, gray symbols represent potential SNPs that do not have sufficient evidence (presence of a new peak with corresponding absence of an expected peak). Larger black symbols indicate a potential SNP with both new peaks and missing expected peaks. Triangles indicate base pair substitution, while circles indicate single base pair deletion.

The data returned from a call to `evaluateSNPs()` includes a list of potential SNPs for each identified novel peak among the input MassArray spectrum. Each novel peak is associated with the following list elements:

1. `SNP` - Contains a list of SNPs, each of which takes the form "`position`:`base`" where `position` is the base pair location within the amplicon sequence, and `base` is the mutated character
2. `SNR` - Contains a numerical list of signal-to-noise ratios corresponding to the expected original peak for the fragment mapping to the identified SNP position
3. `fragment` - Contains a numerical list of fragment IDs which map the SNP position to a specific fragment
4. `SNP.quality` - Contains a numerical list (values ranging from `0` to `2`, with `0` being a highly unlikely SNP and `2` being a SNP with increased likelihood. This number is calculated as a function of new peak SNR and expected peak SNR.
5. `samples` - Contains a list of samples whose spectral data contained the given new peak
6. `count` - Specifies the number of unique SNP and sample pairs, exactly equivalent to the length of `SNP`, `SNR`, `fragment`, `SNP.quality`, or `samples`

Note that each novel peak may be explained by any number of potential SNPs; the list returned only includes the most reliable, but the redundant nature of the data necessitates returning a nested list, as shown below:

```
> length(SNP.results)

[1] 2

> SNP.results[[2]]

$sequence
[1] "AAACGGT"

$fragment
[1] 10  6 10

$SNR
[1] 16.0079 29.8730 29.8730

$SNP
[1] "312:G" "355:T" "312:G"
```

```
$SNP.quality
[1] 1.742775 1.000000 2.000000

$samples
[1] "A" "B" "B"

$count
[1] 3
```

A Previous Release Notes

- No previous releases to date.

References

Sequenom. *EpiTYPER Application Guide*. Sequenom, Inc., San Diego, CA, v1.0.5 edition, 2008.

R.F. Thompson, M. Suzuki, K.W. Lau, and J.M. Greally. A pipeline for the quantitative analysis of cg dinucleotide methylation using mass spectrometry. *Bioinformatics*, 25 (17):2164–2170, 2009.