Genome-wide systematic identification of methyltransferase recognition and modification patterns

Torbjørn Ølshøj Jensen¹, Christian Tellgren-Roth², Stephanie Redl¹,³, Jérôme Maury¹, Simo Abdessamad Baallal Jacobsen¹, Lasse Ebdrup Pedersen¹ & Alex Toftgaard Nielsen¹

Genome-wide analysis of DNA methylation patterns using single molecule real-time DNA sequencing has boosted the number of publicly available methylomes. However, there is a lack of tools coupling methylation patterns and the corresponding methyltransferase genes. Here we demonstrate a high-throughput method for coupling methyltransferases with their respective motifs, using automated cloning and analysing the methyltransferases in vectors carrying a strain-specific cassette containing all potential target sites. To validate the method, we analyse the genomes of the thermophile *Moorella thermoacetica* and the mesophile *Acetobacterium woodii*, two acetogenic bacteria having substantially modified genomes with 12 methylation motifs and a total of 23 methyltransferase genes. Using our method, we characterize the 23 methyltransferases, assign motifs to the respective enzymes and verify activity for 11 of the 12 motifs.
Modification of DNA with a methyl group, catalysed by DNA methyltransferases, is considered to be the most abundant form of post-replicative nucleotide modification found in genomes of both prokaryotes and eukaryotes. In mammalian cells, DNA methylation is known to be involved in a wide range of functions in the regulation of the genome. Methylation of DNA by methyltransferases in bacteria and archaea has mostly been described as part of the sequence-specific restriction modification system (R-M), which traditionally has been viewed as primitive immune systems of bacterial cells to withstand invasion of foreign DNA. However, this conception has been progressively broadened to include additional roles/ functions (recently reviewed by Vasu). Along the R-M system, methylation expands the coding capacity of DNA to include distinction of original from newly synthesized DNA as well as it participates in regulating chromosome replication, nucleoid segregation, conjugation, and global gene expression regulation. Following the nomenclature of R-M systems, which is based on functional arrangement and the requirement for cofactors and/or specific subunits, methyltransferases are classified in three main types (Type I, II, III). In bacteria, methyltransferases typically transfer a methyl group from a donor such as S-adenosyl-L-methionine (SAM) to a target nucleotide base, thereby forming N-6-methyladenine (m6A), N-4-methylcytosine (m4C), or C-5-methylcytosine (m5C). A considerable amount of information regarding both methyltransferases and restriction endonucleases exists, based on laborious experimental approaches such as radioactive labeling with [3H] S-adenosylmethionine, sensitivity to cleavage by restriction endonucleases, and methylation-specific PCR of bisulfite-modified DNA. Recently, genome-wide analyses of methylomes have been published, relying on Single Molecule Real-Time DNA sequencing (SMRT) for direct detection of methylated DNA nucleotides, thereby making it possible to recognize all three types of bacterial methylation. Other sequencing methods, including bisulfite sequencing, Nanopore technology, also have the potential to detect modifications of DNA. SMRT sequencing has even been used to characterize methyltransferases at an individual level. In 2016, Blow et al. assigned motifs to multiple methyltransferases based upon significant similarities to methyltransferases with known specificity and methylome data. With the increasing number of methylomes publicly available, 834 distinct methylation motifs were identified in 230 bacterial and archaeal species and relative easy generation of methylome data (currently still requiring human analysis of the data), the diversity of methylation between species and strains as well as mobility of RM systems or target recognition domain (TRD) has become apparent. Approaches comparing methylomes and genomes to couple TRD of methyltransferases to specific patterns have been published, enabled by a specific degree of diversity and/or size of the compared groups. Deciphering sequence specificity of methyltransferases not depended on information and specific diversity in a group of closely related strains still remains a challenging and laborious task. Algorithms predicting DNA-protein interactions can be applied to facilitate the coupling, but the accuracy of the predictions is typically only around 80%. It is foreseen that experimental evidence could improve these algorithms significantly.

The present study links the DNA methylation pattern identified with SMRT sequencing in a high-throughput manner to sequence-specific enzymatic activities of the assessed methyltransferases. This is achieved by an automated design of strain-specific motif-cassettes containing all potential target sites, cloning of methyltransferases in vectors carrying the cassette, and easy sorting and handling of the data. Analysing methyltransferases from the thermophile Moorella thermaacetica and the mesophile Acetobacterium woodii, two aceticogenic bacteria having substantially modified genomes, demonstrates the applicability and versatility of the method. Using this approach, we characterize 23 methyltransferases, assign motifs to the respective enzymes and verify activity for 11 of the 12 motifs.

Results
Genomic sequence and methylome data. The two aceticogenic strains, M. thermaacetica (thermophilic) and A. woodii (mesophilic), were selected for validating the developed method, since both are distant to the expression host E. coli and have different growth optima. Furthermore, both strains are gas-consuming bacteria with a scarce product spectrum, making them highly interesting hosts for production of biochemicals. However, genetic engineering of the two organisms is still challenging, most likely due to foreign DNA being digested by the native R-M systems.

Initially, the genome and methylome of M. thermaacetica was determined using SMRT sequencing technology, with an average coverage of 137-fold and mean mapped subread length of 11,250 bp using parameters described in the methods section. Previously, similar genome sequencing of A. woodii resulted in an average coverage of 163-fold. The genomes of the strains are relatively small, 2.63 Mb for M. thermaacetica (accession number CP031054) and 4.04 Mb for A. woodii (accession CP002987). The two strains showed significant differences in their methylomes. The thermophilic actogen M. thermaacetica had a high modification frequency of 22 modifications/kb (modifications on either strand), corresponding to every 45th nucleotide on average being modified, whereas A. woodii only had a frequency of 4.4 modifications/kb (Table 1). Modification frequencies of M. thermaacetica is comparable to the most heavily methylated genomes analysed so far by SMRT technology. Average occurrence of three motifs per organisms (in bacteria or archaea) has been observed. In M. thermaacetica seven motifs were identified, five m6A and two m4C modifications, only reads longer than 10 kb were considered and min QV of 85. The motifs included two asymmetric bipartite sequence motifs (characteristic for Type I methyltransferases), these appeared to complement each other, both modified at the first adenine (Table 1). One of the motifs included a “W” as first base, which was not included in the complement motif. The “W” was considered miscalled caused

| Table 1 Summary of methylome data of the two different strains |
| --- |
| Motifs | Modified position | Type | % motifs detected | # of motifs in genome |
| M. thermaacetica 39073-HH |  |  |  |  |
| WATCNNNNNCTC | 2 | m6A | 100.0% | 461 |
| GAGNNNNNGAT | 2 | m6A | 99.4% | 950 |
| SACT | 2 | m6A | 99.9% | 37470 |
| GAWT | 2 | m6A | 99.7% | 4600 |
| AACA | 5 | m6A | 99.7% | 5547 |
| GGGCC | 5 | m4C | 97.4% | 2352 |
| CTCG | 4 | m4C | 76.4% | 5439 |
| A. woodii DSM 1030 |  |  |  |  |
| GCCRAG | 5 | m6A | 99.6% | 1989 |
| CCWG | 1 | m5C | - | 9414 |
| TAAGNNNNNTCC | 3 | m6A | 99.8% | 304 |
| GATNNNNNTGCA | 2 | m6A | 99.8% | 375 |
| CAAAAR | 6 | m6A | 98.9% | 5576 |

The data for A. woodii was from Blow et al.
by the overlapping motif SATC modified similarly at the adenine base. SATC was the most frequently occurring motif in *M. thermoacetica*, occurring 37,470 times in the genome. Furthermore, two non-palindromic motifs were detected, which are typical for Type III and some Type II methyltransferases.

In *A. woodii* five motifs were observed (PacBio data from previous study); four m6A and one m5C modification, none of these were identified in *M. thermoacetica*. The motifs of *A. woodii* included two asymmetric bipartite sequence motifs. The palindromic sequence CCWGG, m5C modified at the first position, was the most frequently occurring motif appearing 9414 times in the genome. The remaining two motifs appeared in total 7565 times, were non-palindromic expressing high specificity with only one degenerate nucleotide in each motif. For *A. woodii*, methyltransferase genes encoded in the genome had been linked to four of the five motifs based on similarities to other already characterized methyltransferases and deduction. Similar bioinformatic analysis of motifs was performed on the genes of *M. thermoacetica*; motifs were predicted for three genes but deviating from the methylome identified by SMRT sequencing. In the two strains, nine modification motifs were assigned to genes using the pipeline at REBASE. However, of the 23 methyltransferase genes more than 60% had unknown target sites. More importantly, a considerable number of modifications identified were not assigned to specific genes. This emphasizes the need for generating more experimental data to support the methyltransferase specificity predictions.

**Motif cassette design and cloning of methyltransferase genes.**

Coupling the methyltransferases to the specific motifs determined by Pacbio-sequencing of the genomes, as described above, would typically require a tedious workflow including cloning, expression, and individual activity determination. The method developed in this study was designed to enable a high-throughput workflow and carry out the sequencing in a single library without barcoding (Fig. 1). Central to the high-throughput workflow was the design of the methylome-specific motif cassette. The motif cassettes included all motifs randomly organized, assuring that copies of the same motifs were not placed directly adjacent to each other. Motifs including unspecific regions of 2–10 n’s (typical for Type I motifs) is present 12 times in the cassette, whereas motifs not including such regions were repeated 10 times.

![Fig. 1 MetMap workflow. The box designated “automated” indicates the steps that were automated. The variance in the occurrence of motifs in the motif-cassette depends on the nature of the motifs; A motif including unspecific regions of 2–10 n’s (typical for Type I motifs) is present 12 times in the cassette, whereas motifs not including such regions is repeated 10 times.](image)

SMRT-seq of pooled plasmids expressing single methyltransferases. Eleven and fifteen different plasmids with individual methyltransferases were pooled and sequenced using one SMRT-cell (a pool of up to 19 plasmids were sequenced in one SMRT-cell with sufficient coverage). Sequences of the individual
methyltransferases and negative controls were isolated using a two-step sorting process (outlined in Fig. 2): initial sorting based only on the strain-specific motif cassette, followed by a sorting where the reads were clustered according to the methyltransferase sequences and the motif cassette. Thereafter, the modification patterns of each single methyltransferase were determined.

Of the 23 methyltransferases and two negative controls, it was possible to isolate all coding sequences and identify 21 modification motifs with a detection rate close to 100% for all (Supplementary Data 1). Additional motifs were detected (Supplementary Data 1 lists all information from the motif_summary.csv files) under the given parameters, however upon inspection some appeared as miscalled motifs. The coverage ranged from 116- to 11,339-fold (using std. parameters), and we found no correlation between coverage and the DNA concentration or size (see Supplementary Fig. 1 and Supplementary Fig. 2). The large coverage span did not make it possible to analyze all motifs identifications with the speciﬁc subunit located close to the gene). Two sets (Awo_c04420_Awo_c04450 and Awo_c17130_Awo_c17140) were expected. Motif identiﬁcation of Awo_c17130 (DNA speciﬁcity subunit) in the genome this may be more challenging). The only modiﬁcations identiﬁed, have motifs corresponding to the motifs for each sample the minimum read score was adjusted to reduce the coverage and increase the modiﬁcation QV cut-off to avoid false positive identiﬁcations of highly degenerate motifs.

Detection of motifs from different methyltransferase types. To demonstrate the versatility of the method, methyltransferases annotated as both Type I, II, and III were included. One thermophilic Type I methyltransferase, encoded by the two genes MothHH_01799 (methyltransferase subunit) and MothHH_01798 (DNA speciﬁcity subunit), was included in the analysis pipeline. Detection of the methylation signal requires different read coverage depending on the modiﬁed base. Detection of m5C requires at least 5- to 10-times more coverage than for m6A and m4C, due to lower kinetic variation signals. Therefore for each sample the minimum read score was adjusted to reduce the coverage and increase the modiﬁcation QV cut-off to avoid false positive identiﬁcations of highly degenerate motifs.

The identiﬁed modiﬁcation motifs with the speciﬁc methyltransferase genes are summarized in Tables 2 and 3. Eleven of the methyltransferases to motifs observed in the genomes. Awo_c17840 did not show any activity, which could be explained by the lack of a negative controls. The expressed control genes were assessed for their functionality in the presence of 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (x-gal).
speciﬁcity subunit. REBASE predicted motifs for both Type I methyltransferases of A. woodii (Awo_c08800_Awo_c08810 and Awo_c17130_Awo_c17140). The prediction was based on similarities of Awo_c08800 to S.Kaq16071III and deduction (excluding potential nonfunctional methyltransferases). Our results experimentally validate the predictions, and demonstrate the strength of this method to detect modification by Type I methyltransferases directly.

**Table 2 Summary of the methyltransferases from M. thermoacetica**

| Gene name ATCC 39073-HH | Gene name ATCC 39073-HH | Gene product | Predicted speciﬁcity | Motif detected | Mean mod. QV | Coverage |
|-------------------------|-------------------------|---------------|----------------------|----------------|-------------|----------|
| Type I                  |                         |               |                      |                |             |          |
| MothHH_01798            | Moth_1671               | DNA met. Speciﬁcity | N.A.                | WGA(m6A)*GNNNNAGATNNNNC GNNBNA(m6A)*TCTNNNWNTC | 55.40 | 1643.71 |
| MothHH_01799            | Moth_1672               | N-6 DNA methylase | N.A.                |                | 53.80 |          |
| Type II                 |                         |               |                      |                |             |          |
| MothHH_00029            | Moth_0026               | DNA adenine methylase | GASTC | GA(m6A)*WTC BGNSTSC(U.K.) NNGNNNNTGC | 49.58 | 50.29 |
| MothHH_00692            | Moth_0639               | DNA methylase N-4/N-6 | N.A. | GHA(m6A)*NCCC VNGNNNNTGC | 35.40 | 224.54 |
| MothHH_00927            | Moth_0871               | conserved hypothetical protein | N.A. |                | 88.75 | 95.23 |
| MothHH_01002            | Moth_1737               | DNA adenine methylase | GGATC | GA(m6A)*TC N.D. N.D. N.D. | 53.14 | 60.88 |
| MothHH_02421            | Moth_2238               | hypothetical protein | N.A. | N.D. | 6388 |          |
| MothHH_02423            | Moth_2241               | Adenine-speciﬁc DNA methylase | AGGCCT | GGA(m6A)*C NGNNNNTGC | 1136 | 1768 |
| Type III                |                         |               |                      |                |             |          |
| MothHH_01297            | Moth_1213               | Site speciﬁc subunit | N.A. | AACCA(m6A) | 69.93 | 83.43 |
| Neg-control             | bgaB (beta-galactosidase) | None | N.D. | N.D. | 116 |          |

Methyltransferases from M. thermoacetica cloned individually and their corresponding motifs
N. A. not available, N. D. not detected, U. K. unknown
*indicates that the methylation type was identiﬁed by the modiﬁcation position

**Table 3 Summary of the methyltransferases from A. woodii**

| Gene name (DSMZ 1030) | Gene product | Predicted speciﬁcity | Motif detected | Mean mod. QV | Coverage |
|-----------------------|--------------|----------------------|----------------|-------------|----------|
| Type I                |              |                      |                |             |          |
| Awo_c04420            | R-M subunit HsdM1 | N.A. | N.D. | N.D. | 2075 |
| Awo_c04450            | R-M subunit HsdS1 | N.A. | N.D. | N.D. | 2075 |
| Awo_c08800            | R-M subunit HsdM2 | TAAGNNNNNTCC | TAA(m6A)*GNNNNNTCC | 1914 | 3446 |
| Awo_c08810            | R-M subunit HsdS2 | GATGNNNNNNTGC | GA(m6A)*NTNNTGCC | 1136 | 1768 |
| Awo_c17130            | R-M subunit HsdM3 | N.A. | N.D. | N.D. | 1515 |
| Awo_c17140            | R-M subunit HsdM4 | N.A. | N.D. | N.D. | 1471 |
| Type II               |              |                      |                |             |          |
| Awo_c06460            | DNA (cytosine-5-) -methyltransferase Dcm | CCGG | N.D. | N.D. | 8569 |
| Awo_c14460            | DNA modiﬁcation methyltransferase | N.A. | CAAAAA(m6A)*R | 2185 | 6372 |
| Awo_c18590            | Putative DNA methylase | N.A. | TNNNCTG(U.K.)*GNNNNV MNNBSNNTNCNG(U.K.)*G | 48 | 8514 |
| Awo_c19740            | Potential Type II | GCRAG | N.D. | N.D. | 6223 |
| Awo_c30860            | Putative DNA methylase | N.A. | TA(m6A)*GNHNNTNV BCANA(m6A)*NGNNNNTNM VTNNNNNTTNTA(m6A) | 608 | 5539 |
| Awo_c31130            | DNA (Cytosine-5-) -methyltransferase YdP | N.A. | N.D. | N.D. | 11339 |
| Awo_c34920            | N6 adenine-speciﬁc DNA methylase phage N6 methylase | N.A. | WBNNMVDVA(m6A)*GCNH WNNNNMGNM(m4C) | 70 | 937 |
| Awo_c35180            | CCWGG | N.D. | 237 | 708 |
| Awo_c35200            | Lambda C methyl | N.A. | N.D. | N.D. | 909 |
| Neg-control           | bgaB (beta-galactosidase) | None | N.D. | N.D. | 8051 |

Methyltransferases from A. woodii cloned individually and their corresponding motifs
*indicates the modiﬁcation position is misleading, as the modiﬁcation on a guanine is not likely, in addition m5C has typical characteristic kinetic signals two and six bases downstream of the methylated position
N. A. not available, N. D. not detected, U. K. unknown
Of the seven Type II methyltransferases from *M. thermoacetica*, it was possible to assign motifs to four of them. Interestingly, the motif SATC identified on the genome turned out to be a double motif of GATC and CATC, modified by two different methyltransferases MothHH_01869 and MothHH_00927, respectively. The motif CTCGG was not properly identified by any of the cloned methyltransferases or the tested combination of TRD and methyltransferases, given the low detection rate on the genome sequencing (Table 1) this was expected. However, the motifs of MothHH_00692 indicated that it could have activity towards the CTCGG motif (extended motif including the motifs STSCN). Six of the ten cloned *A. woodii* methyltransferases displayed activity, three of which corresponded to the modification on the genome, and the remaining three were assigned to promiscuous motifs. Activity of Awo_c18590 resulted in recognition of four motifs, all with relative low modification QV (32–52), associating the motifs to the methylome, the activity could be assigned to motif CCWGG. Despite a coverage of 8514 (>250) modification QV was low and the motif extended, this indicates that detection of m5C by PacBio is incomplete. The motif for Awo_c19740 was identified as GCGGRC/CC/AGC34 correlating with the predicted motif at REBASE (57% identity at protein level to atcII from *Aeroucoccus tetradius*). Furthermore, we identified the methyltransferases Awo_c14460 to be responsible for the CAAAAR modification. Awo_c14460 did not have any predicted specificity. Awo_c30860, Awo_c34920, Awo_c35180, and M00910 were found to have unspecific activity, which could be because the target motifs might not have been included in the cassette, and were thus not present in sufficient number to determine the specific motifs. It could also be because the TRD or specificity subunit was not expressed simultaneously. It was observed that two of the genes Awo_c34920 and Awo_c35180 are encoded on prophage regions, and the genes might therefore not be active in the genome as prophage-encoded genes generally are transcriptionally inactive until the prophage is excised46. *M. thermoacetica* had one annotated Type III methyltransferase, its expression lead to the detection of a non-palindromic motif AACCA. No Type III methyltransferases were annotated in the genome of *A. woodii*. Despite different modes of action, both Type II and Type III methyltransferases can recognize non-palindromic motifs. This demonstrates that the developed method is also suitable for Type III methyltransferases.

Restriction endonuclease digestion of modified plasmids. In order to verify the modification motifs identified with the presented workflow, selected modifications were analysed by restriction digestion. The motifs that could be verified by digestion with commercial available restriction enzymes included GGGGC, GATC, GAWTC, and CCWGG modified by the methyltransferases MothHH_02467, MothHH_01869, MothHH_00029, and Awo_c18590, respectively. Using the restriction enzymes ApaI, DpnI, PfIeI, and EcoRII, all modifications were verified by hindrance or facilitation (DpnI) of digestion (Supplementary Fig. 3 and Supplementary Table 1).

Comparison of predicted and detected motifs. The sequence specificity of the methyltransferases predicted at REBASE is based on nucleotide sequence and similarities to characterized methyltransferases. Generally, the accuracy (percentage of correct prediction) of methods predicting protein-DNA binding sites are increasing but are usually about 70–90%36. By generating more experimental evidence, the specificity and accuracy of these models is expected to significantly increase. When comparing the motifs predicted by REBASE and those found experimentally in this study, some inconsistencies were observed (see Tables 2 and 3). In total, nine methyltransferases had assigned predicted specificity (REBASE), however only in four cases (all from *A. woodii*) were there a correlation with the specificity detected in this study. REBASE was able to predict three motifs for the methyltransferases of *M. thermoacetica*; GASTC, GGATC, and AGGCCT for MothHH_00029, MothHH_01869, and MothHH_02467, respectively. The present study determines the motifs to be GAWTC, GATC, and GGGCCC. Six methyltransferases from *A. woodii* had motifs assigned30 (bioinformatic). Four of these correspond to four of the five motifs found in the methylome. In this study, the activity of the four methyltransferases was demonstrated and the methyltransferase responsible for the fifth motif, CAAAAR was also identified.

Discussion

In this study we presented MetMap, a novel method/pipeline to experimentally demonstrate the DNA modification specificity of methyltransferases in a high-throughput manner (summarized in Fig. 1). A total of 23 methyltransferases and a negative control (beta-galactosidase) were analyzed. Methylase activity was observed from 14 of the methyltransferases, and of 11 of these specificity could be related to their native activity on the genomes. The high throughput and scalability was facilitated by conducting the cloning steps using an automated cloning platform, but relying on gene amplification from genomic DNA. The advancements in gene synthesis methods have enabled assembly of large DNA fragments, with high fidelity at low costs. Exploiting newer lower-cost synthesis methods is expected to bring the cost further down47. In the future, such technology could effortlessly be implemented in the developed methods, and increase the throughput of our method even more.

The evaluated methyltransferases were derived from two different gram-positive acetogenic strains that are distant from the expression host *E. coli*, both with respect to growth temperature and codon usage. These strains were chosen specifically to demonstrate the robustness of the developed method. The methyltransferases were all expressed and evaluated without prior optimization, neither genes nor conditions. Heterologous expression of thermophilic enzymes in mesophilic systems has been published, however, this often requires either additional cofactors, appropriate posttranslational processing, or specific components during multisubunit assembly to enable production in an active form48–52. In this study, expression of the thermophilic methyltransferases at lower temperatures did not affect the activity of the proteins to a degree where it hampered the detection of the respective motifs. In fact, the frequency at which modification motifs could be linked to genes was even higher for methyltransferases derived from the thermophilic strain. Others have reported successful heterologous expression of thermostable methyltransferases in mesophilic systems52,53, however it must be stressed that this observed compatibility of thermostable methyltransferases may not be general. Activity detection of the thermophilic Type I methyltransferases (multiple component enzyme) was arduous, as modQV was low and the motif slightly extended. This was in contrast to the high modQV and accurate motifs when assessing the Type I methyltransferases from the mesophilic host. These differences indicate that the thermophilic holoenzyme may have had reduced activity of reasons elaborated above, however, the method still worked. Maintaining a constant temperature (37 °C) during expression simplifies the experimental pipeline. Furthermore, it prevents induction of the host cells’ heat shock response, which could potentially result in increased expression of undesired proteins such as proteases54. Utilizing an expression host with a better matching codon usage could potentially increase the protein synthesis rate and thereby result in higher methyltransferase activity55. Chemical gene
synthesis of the methyltransferases would allow for individual codon optimization, which may also help to increase the translation of the methyltransferases.

In the present study, we were able to analyze a pool of up to 19 linearized plasmids in a single SMRT cell while reaching average coverage of 4750-fold. This indicates that even a greater number of methyltransferases could potentially be analyzed in one sequencing cell. However, some methyltransferases stand out by having low coverage e.g. the negative control with the M. thermoacetica motif cassette and different modification types require different coverage. Therefore, increasing the number of plasmids pooled for sequencing beyond 20 is not recommended if high detection rates are anticipated. By using larger sized chips the number of plasmids pooled could be further increased. The chip on the Sequel System of PacBio is predicted to handle a pool of 200–300 plasmids, because of improved read-length and loading. In this study m3C methylation was detected by PacBio, but with low accuracy despite a coverage higher than 250-fold. Improvement to accuracy and sensitivity when detecting m3C modifications can be achieved by oxidation of the modified cytosine or by utilizing another sequencing technology.

The application of SMRT and nanopore sequencing has empowered the identification of DNA modifications in a much faster manner than before. This has facilitated better understanding of methylation in cell physiology (mostly based upon sequencing of entire genomes). However, detecting patterns of single methyltransferases remain a challenging task. Heterologous expression of methyltransferases (from plasmid) and subsequently assessing the methylation patterns by SMRT sequencing of both plasmids (various plasmids are needed as methylation sites needs to be present) and genomes of the expressing hosts has been the method of choice. Modification by all three major types of methyltransferases (Type I, II, and III) has been determined, and clearly demonstrates the applicability of SMRT sequencing to decipher the recognition sites of methyltransferases. Compared to the previous time-consuming methods based upon restriction digestion the gain has been massive. The presented method shares main principles of determining modified bases with the previous studies, but targets a more streamlined high throughput approach for both cloning and analysis. This is facilitated by using the same expression vector with a strain specific motif cassette for the methyltransferases, enabling identification of motifs solely based upon the cassette. This was demonstrated by including Type I methyltransferases (two target sequences separated by typical 4–8 nucleotides), which requires a large sequence space to obtain the correct motifs. The two mesophilic Type I methyltransferases were assigned with high specificity to motifs with modQV above 1,000, whereas the thermophilic had slightly extended motifs and lower ModQV (causes are discussed above), but still possible to identify the activity and correlate it to the methylome. Methyltransferases with no activity detected could potentially be functional but not active on the genome. Such sites could potentially be further analyzed by PacBio sequencing the genome of the E. coli strains expressing the individual methyltransferases, thereby obtaining higher specificity in the motif-recognition.

Furuta et al. assigned specificity to methyltransferases or to their specificity-determining domains (TRD) by assessing different strains of Helicobacter pylori for differences in the repertoire of methylation motifs and in specificity-determining genes. A commensurate approach of sequencing closely related strains was done by Chen et al., additionally they combined it with prediction of sequence specificity using SEQWARE. Such a combinatorial approach holds great potential, but sample space has some requirement as it need a specific diversity assuring differences in methylome along with minor changes in the genomes (preferable only in the TRD). The developed method does not have any requirements for a specific diversity on methylome or DNA level. It is anticipated that the data generated with the presented method could contribute in increasing the specificity of the combinatorial methods.

The two strains M. thermaoacetica and A. woodii had both substantially modified genomes. Using the developed method MetMap, we were able to couple the genetic sequences to the methylome patterns in a high-throughput manner. Of the 12 motifs identified by PacBio sequencing of the genome, we were able to directly assign 11 methyltransferases, corresponding to an efficiency of 92%. With the currently available predictive tools (REBASE) nine of the evaluated methyltransferases had assigned specificity, however only in four cases (all from A. woodii) were there a correlation with the specificity detected in this study. By analyzing larger numbers of methyltransferases, the experimentally validated knowledge gained could undoubtedly improve the accuracy of algorithms predicting protein-DNA binding sites based on the protein coding sequence.

Methods

 Plasmid cultivation. The strains and plasmids used in this study are listed in Supplementary Data 2 and the DNA oligomers used for the constructs are listed in Supplementary Data 3. E. coli Top10 (Thermo Fisher Scientific) was used for plasmid constructions, whereas E. coli Stellar® dcm-‘dam’ (Clonetech) was used for expressing the methyltransferases. Both E. coli strains were routinely cultivated in LB (Luria-Bertani) broth at 37 °C with shaking. For plasmid propagation TB (Terrific Broth) was used. When required respective antibiotic was supplemented to the media. M. thermoacetica ATCC 39073-HH was grown at 60 °C in undefined media (Supplementary Table 2, in accordance with Daniel et al.) utilizing fructose as carbon source. A. woodii DSMZ 1030 was grown in DSMZ medium 135 strictly anaerobically at 30 °C.

 Genome extraction. Cultures of M. thermoacetica and A. woodii were grown until exponential growth was reached. Then gDNA was extracted by Wizard® Genomic DNA Purification Kit (Promega), according to manufactures guidelines, with the following modifications. The lysis step with 10 mg/ml lysosome was extended to 2 h and gDNA was resuspended in 1 mM Tris-HCL (pH 7.4) and 0.1 mM EDTA (pH 8.0). DNA quantification was done fluorometrically utilizing a Qubit® instrument (Thermo Fisher Scientific).

 Genome sequencing. M. thermoacetica was sequenced using a PacBio RSII instrument (Pacific Biosciences). DNA was sheared into 10 kb fragments using a Genemachines HydroShear Instrument (Digilab). SMRT-bells were constructed according to the manufacturer’s instructions (Pacific Biosciences). SMRT-bells were sequenced on two SMRT-cells on a Pacific Biosciences RSII sequencer according to the manufacturer’s instructions with 4 h movie-time. Only reads longer than 10 kb were considered and min QV of 85 considered. The average reference coverage was above x137, from 49,792 reads, with an average read length of ~11,000 bp.

 Program for designing motif cassettes. A script was designed to randomly organize the motifs initially identified by the PacBio sequencing of the genomes and stitch them together in adequate number, resulting in the strain-specific motif cassette. Furthermore, the script separates each motif by a ‘n’. The user submits the motifs to the program, classified by two rules. Rule 1: motifs of 2–10 nt’s separated by two specific sites, typical for Type I methyltransferases. Rule 2: motifs not having separated specificity sites, generally for Type II and III. This differentiation is included to assure higher number of motifs for Type I motifs (12 vs. 10 motifs). For motifs with ambiguous nucleotides, are de-ambiguated, according to two rules: Rule 1, pick L random variants of the motif. As an example, the motif ATGNNTTTA has a total of 16 possible actual sequences. If L ≤ 16 then the program will randomly pick L variants (without duplicates). If L >16 then each possible variant will be picked at least L/16 times and some will be picked 1 more than that. Rule 2, make K copies of each completely “de-ambiguated” variant: E.g. the sequence “SATC” will then be treated as two sequences: “GATC” and “CTAC” that will each appear in K copies. The program outputs optional number of motif cassettes as individual genbank and fasta files. The script for designing the motif cassettes was developed to run in python 3.6 or newer and can be found together with program and installation descriptions at https://github.com/biosustain/metmap.

 Construction of plasmids. Plasmids (listed in Supplementary Data 2) were constructed by PCR-based uracil-specific excision reagent (USER) cloning method or using Gibson cloning. For USER cloning, one microliter of 5x HF buffer
(Thermo Scientific) and 1U of USER® enzyme mix (New England Biolabs, 1 U/µl) were added to 8 µl of the mixture of purified PCR products, plasmid-backbone or genes. All PCR products were amplified with oligonucleotides (Integrated DNA Technologies) having uracil incorporated, utilizing Phusion U polymerase (Thermo Fisher Scientific). The reaction mixture was incubated for 25 min at 37 °C, followed by 25 min of incubation at a temperature optimized for annealing of the fragments for 25 min. Eight microliters MillQ was added to the reactions, reaching a final concentration of 1 µM. In all, 2.5 µl of the diluted USER reaction was used to transform 50 µl competent E. coli. Gibson assemblies were carried out using Gibson Assembly Cloning Kit (NEB). Fragments (motif cassette) with ~30 bp overlap regions were mixed together with the backbone fragment in 1:1 molar ratio together with an equal volume of Gibson master mix, incubated at 50 °C for 1 h and immediately transformed into E. coli.

Automated cloning. An automated cloning workflow was set up on a Hamilton Microlab VANTAGE Liquid Handling System (Hamilton Bonaduz AG). The system was fully customized to accommodate complex strain construction workflows. It was equipped with a modular arm holding eight single pipetting channels (Single Channel 1000-µl NanoPulse), a 96-channel pipetting head (MPH, Multiprobe Head), a camera, a barcode reader (ID Loading Device (1D & 2D Barcode Reading)), four heat shakers (Hamilton), a vacuum station CVS (Hamilton), a HEPA hood (Hamilton), a Microplate centrifuge and automated Microplate centrifuge loader (Agilent Technologies Inc), and a TRobot thermocycler (Analytik Jena). Cloning was using in general conditions described above. Analysis of DNA fragments by high throughput capillary electrophoresis was performed off-a-deck on a Caliper LabChip GXII (PerkinElmer Inc.).

The liquid handling system was loaded with genomic DNA from the different organisms, oligonucleotides (Supplementary Data 3) for amplifying the 23 methyltransferases and a negative control (the bgfd gene from Geobacillus stearothermophilus, refseq WP_020755758.1), and plasmid backbone holding a strain-specific motif cassette. Clones were verified by colony PCR and subsequent Sanger sequencing of the insert. To reach complete cloning efficiency, the automated cloning was complemented by manually cloning, allowing for individual optimization of selected PCR conditions.

Plasmid extraction. Fresh medium (supplemented with antibiotic) was inoculated with 10% overnight culture. Growth was re-established by incubation at 37 °C for 1 h before induction with IPTG (final concentration of 1 mM). The induced cultures were incubated for additional 3 h. The plasmids were purified using a Nucleospin® Plasmid (Macherey-Nagel) spin column kit. DNA was quantified fluorometrically utilizing Ovation® (Thermo Fisher Scientific). Prior to pooling, the plasmids were analyzed with NotI (Thermo Fisher Scientific) and the enzyme was inactivated by incubation at 80 °C for 10 min.

SMRT sequencing of pooled plasmids. A SMRT-bell library was constructed for each pool of linearized plasmids according to the manufacturer’s instructions (Pacific Biosciences). SMRT-bells were sequenced on one SMRT-cell per pool on a Pacific Biosciences RSII sequencer according to the manufacturer’s instructions with 4h movie-time.

Bioinformatics/sequence handling. Analysis of the methylation profile and motif detection was performed in SMRT-Portal v.2.3.0 as follows: Reads for the individual methyltransferase sequences and negative controls were separated in two steps: an initial sorting based on the motif cassette was followed by grouping based on the specific methylene sequence. The methylation status of the selected sequencing reads was then determined using the ‘Modification and Motif_Analysis’ pipeline included in SMRT-Portal (details are available at https://www.pacb.com/wp-content/uploads/2015/09/SMRT-Pipe-Reference-Guide.pdf). When motif identification by standard parameters was too ambiguous and or appeared extended, optimization was conducted to assure better motif identification. In particular, measures were taken to reduce coverage into the by the manufactures reported coverage range, including adjustment of min read length, readQV, and ModQV.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated and analyzed during the current study are available online, including the newly annotated genome of M. thermotolerans 39073-H1H (Genbank accession number CP031054) and SMRT sequencing files SRR8538415, and SRR8537235 Motif_summary.csw files are supplied as Supplementary Data 1.

The code developed for designing the motif cassettes can be found together with program and installation description at github (https://github.com/biosustain/metmap).

Received: 19 November 2018 Accepted: 27 June 2019
Published online: 19 August 2019

References

1. Wion, D. & Casadesús, J. N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. Nat. Rev. Microbiol. 4, 183–192 (2006).
2. Clark, T. A. et al. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Res. 40, e29 (2012).
3. Jones, P. A. & Takai, D. The role of DNA methylation in mammalian epigenetics. Science 293, 1068–1070 (2001).
4. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. Trends Biochem. Sci. 31, 89–97 (2006).
5. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. Nat. Rev. Genet. 14, 204–220 (2013).
6. Pingoud, A., Wilson, G. G. & Wende, W. Type II restriction endonucleases—a historical perspective and more. Nucleic Acids Res. 42, 7489–7527 (2014).
7. Pingoud, A., Fuxreiter, M., Pingoud, V. & Wende, W. Type II restriction endonucleases: structure and mechanism. Cell. Mol. Life Sci. 62, 685–707 (2005).
8. Loenen, W. A., Dryden, D. T., Raleigh, E. A. & Wilson, G. G. Type I restriction enzymes and their relatives. Nucleic Acids Res. 42, 20–44 (2014).
9. Rao, D. N., Dryden, D. T. & Bheemanaik, S. Type III restriction-modification enzymes: a historical perspective. Nucleic Acids Res. 42, 45–55 (2014).
10. Loenen, W. A., Dryden, D. T., Raleigh, E. A., Wilson, G. G. & Murray, N. E. Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Res. 42, 3–19 (2014).
11. Sánchez-Romero, M. A., Cota, I. & Casadesús, J. DNA methylation in bacteria: from the methyl group to the methylene. Curr. Opin. Microbiol. 25, 9–16 (2015).
12. Vasu, K. & Nagaraja, V. Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol. Mol. Biol. Rev. 77, 53–72 (2013).
13. Camacho, E. M. & Casadesús, J. Regulation of trh transcription in the Salmonella virulence plasmid by strand-specific DNA adenine hemimethylation. Mol. Microbiol. 57, 1700–1718 (2005).
14. Roberts, D., Hoopes, B. C., McClure, W. R. & Kleckner, N. IS10 transposition in E. coli methylase sequence. The methylation status of the selected
15. Srikhanta, Y. N., Maguire, T. L., Stacey, K. J., Grimmond, S. M. & Jennings, M. P. The phasevarion: a genetic system controlling coordinated, random switching of expression of multiple genes. Proc. Natl Acad. Sci. USA 102, 5547–5551 (2005).
16. De Bolle, X. et al. The length of a tetranucleotide repeat tract in Haemophilus influenzae determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol. Microbiol. 35, 211–222 (2000).
17. Wilson, G. G. & Murray, N. E. Restriction and modification systems. Annu. Rev. Genet. 25, 585–627 (1991).
18. Roberts, R. J. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res. 31, 1805–1812 (2003).
19. Cheng, X. Structure and function of DNA methyltransferases. Annu. Rev. Biophys. Biomol. Struct. 28, 293–318 (1995).
20. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE - a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 38, D234–D236 (2010).
21. Bitinaite, J. et al. Alw26l, Eco311 and Esp31—type IIs methyltransferases modifying cytosine and adenine in complementary strands of the target DNA. Nucleic Acids Res. 20, 4981–4985 (1992).
22. Singersam, J. et al. Use of a HpaII-polymerase chain-reaction assay to study DNA methylation in the Pkg-1 Cpg island of mouse embryos at the time of X-chromosome inactivation. Mol. Cell. Biol. 10, 4987–4989 (1990).
23. Herman, J. G. et al. A novel PCR assay for methylation status of Cpg islands. Proc. Natl Acad. Sci. USA 93, 9821–9826 (1996).
24. Plulberg, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods 7, 461–465 (2010).
25. Chao, M. C. et al. A cytosine methyltransferase modulates the cell envelope stress response in the Cholera pathogen [corrected]. PLoS Genet 11, e005666 (2015).
26. Khabramanoglou, C. et al. Genomics of DNA cytosine methylation in Esherichia coli reveals its role in stationary phase transcription. Nat. Commun. 3, 866 https://doi.org/10.1038/ncomms1878 (2012).
27. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods 14, 407–410 (2017).
28. Liu, L. et al. Recent advances in the genomic profiling of bacterial epigenetic modifications. Biotechnol. J. 14, e1800001 (2019).
Kandror, O., Busconi, L., Sherman, M. & Goldberg, A. L. Rapid degradation of

Chung, D., Farkas, J., Huddleston, J. R., Olivar, E. & Westpheling, J.

NATURE COMMUNICATIONS | (2019) 10:3311 | https://doi.org/10.1038/s41467-019-11179-9 | www.nature.com/naturecommunications

Kurland, C. & Gallant, J. Errors of heterologous protein expression. Nat. Commun. 10, 10.1038/s41467-019-11179-9 (2019).

Kandror, O., Busconi, L., Sherman, M. & Goldberg, A. L. Rapid degradation of DNA methyltransferase sequence specificity. PLoS Genet. 10, e1004272 (2014).}

Krebbs, J. et al. The complex methylome of the human gastric pathogen Helicobacter pylori. Nucleic Acids Res. 42, 2415–2432 (2014).}

Chen, P. et al. Comparative genomics reveals the diversity of restriction-modification systems and DNA methylation sites in Listeria monocytogenes. Appl. Environ. Microbiol. 83, https://doi.org/10.1128/AEM.02091-16 (2017).

Si, J., Zhao, R. & Wu, R. An overview of the prediction of protein DNA-binding sites. Int. J. Mol. Sci. 16, 5194–5215 (2015).

Poehlein, A. et al. An ancient pathway combining carbon dioxide fixation with the generation and utilization of a sodium ion gradient for ATP synthesis. PLoS ONE 7, e32909 (2012).

Fang, G. et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nat. Biotechnol. 30, 1232–1239 (2012).

Murray, I. A. et al. The methylomes of six bacteria. Nucleic Acids Res. 40, 11450–11462 (2012).

Furuta, Y. et al. Methylome diversification through changes in DNA methyltransferase sequence specificity. PLoS Genet. 10, e1004272 (2014).}

Pacific Biosciences. White paper, Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing. Available at: https://www.pacb.com/wp-content/uploads/2015/09/VP_Detecting_DNA_Base_Modifications_U.S_.pdf (accessed: 14th September 2017)

Beaulaurier, J. et al. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. Nat. Commun. 6, 7438 (2015).

Rhoods, A. & Ku, K. PacBio sequencing and its applications. Genom. Proteom. Bioinforma. 13, 278–289 (2015).

Jensen, T. O., Pogrebnyakov, I., Falkenberg, K. B., Redl, S. & Nielsen, A. T. Application of the thermotolerant beta-galactosidase, BgaB, from Geobacillus stearothermophilus as a versatile reporter under anaerobic and aerobic conditions. AMB Express 7, https://doi.org/10.1186/s13568-017-0469-2 (2017).

Suzuki, H., Murakami, A. & Yoshida, K. Counterselection system for Geobacillus kaustophilus HTA426 through disruption of pyrF and pyrR. Appl. Environ. Microbiol. 78, 7376–7383 (2012).

Ventura, M. et al. Comparative genomics and transcriptional analysis of prophages identified in the genomes of Lactobacillus gasseri, Lactobacillus salivarius, and Lactobacillus casei. Appl. Environ. Microbiol. 72, 3130–3146 (2006).

Hughes, R. A. & Ellington, A. D. Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. Cold Spring Harb. Perspect. Biol. 9, https://doi.org/10.1101/cshperspect.a028127 (2017).

Moreno, R., Haro, A., Castellanos, A. & Berenguer, J. High-level overproduction of His-tagged Tth DNA polymerase in Thermus thermophilus. Appl. Environ. Microbiol. 71, 591–593 (2005).

Koma, D., Sawai, T., Harayama, S. & Kino, K. Overexpression of the genes from thermophiles in Escherichia coli by high-temperature cultivation. Appl. Microbiol. Biotechnol. 73, 172–180 (2006).

Hess, M. Thermoadapohilic proteins for biofuel production. Trends Microbiol. 16, 414–419 (2008).

Hidalgo, A. et al. Thermus thermophilus as a cell factory for the production of a thermophilic Mn-dependent catalase which fails to be synthesized in an active form in Escherichia coli. Appl. Environ. Microbiol. 70, 3839–3844 (2004).

Kreftl, D., Papkov, A., Zylcz-Stachula, A. & Skowron, P. M. Thermostable proteins bioprocesses: the activity of restriction endonuclease-methyltransferase from Thermus thermophilus (RM.TthHB27) cloned in Escherichia coli is critically affected by the codon composition of the synthetic gene. PLoS ONE 12, e0186633 (2017).

Chung, D., Farkas, J., Huddleston, J. R., Olivar, E. & Westpheling, J. Methylation by a unique alpha-class N4-cytosine methyltransferase is required for DNA transformation of Caldibacillus solitariorum besci DSM6725. PLoS ONE 7, e43844 (2012).

Kandror, O., Busconi, L., Sherman, M. & Goldberg, A. L. Rapid degradation of an abnormal protein in Escherichia coli involves the chaperones GroEL and GroES. J. Biol. Chem. 269, 23575–23582 (1994).

Kurland, C. & Gallant, J. Errors of heterologous protein expression. Curr. Opin. Biotechnol. 7, 489–493 (1996).

Clark, T. A. et al. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. BMC Biol. 11, https://doi.org/10.1186/1741-7007-11-4 (2013).