MeStudio: crossing methylation and genomic features for comparative epigenomic analyses

Christopher Riccardi[1], Iacopo Passeri[1], Lisa Cangioli[1], Camilla Fagorzi[1], Alessio Mengoni[1], and Marco Fondi[1]

[1]University of Florence

June 10, 2022

Abstract

DNA methylation is one of the most relevant epigenetic modifications. It is present in eukaryotes and prokaryotes and is related to several biological phenomena, including gene flow and adaptation to environmental conditions. The widespread use of third-generation sequencing technologies allows direct and easy detection of genome-wide methylation profiles, offering increasing opportunities to understand and exploit the epigenomics landscape of individuals and populations. Here, we present MeStudio, a pipeline which allows to analyse and combine genome-wide methylation profiles with genomic features. Outputs report the presence of DNA methylation in coding sequences (CDS) and noncoding sequences, including both intergenic sequences, and sequences upstream to CDS. We show the usage and performances of MeStudio on a set of single-molecule real time sequencing outputs from strains of the bacterial species Sinorhizobium meliloti. MeStudio is freely available under an open source GPLv3 license at https://github.com/combogenomics/MeStudio

1 Introduction

Understanding organism adaptation to variable environmental conditions is pivotal for weighting the relevance of natural selection over species and population evolution. Phenotypic plasticity, stress responses and acclimation display significant contribution from epigenetic mechanisms (Moler *et al.* , 2019). Among epigenetic modifications, DNA methylation has been shown to be key in the control of several biological phenomena in eukaryotes and prokaryotes (Jones, 2012) and in the last years the study of variation in epigenetic response is stirring the attention of several investigators (Chen *et al.* , 2020). Third-generation sequencing technologies, namely single molecule real-time (SMRT) (Flusberg *et al.* , 2010; Fang *et al.* , 2012) and nanopore ONT (Clarke *et al.* , 2009; Simpson *et al.* , 2017) sequencing allow to directly identify the most commonly methylated bases (Gouil and Keniry, 2019; Sánchez-Romero and Casadesús, 2020; Rand *et al.* , 2017). These methods are boosting genome-wide DNA methylation studies, especially in prokaryotes, where the compact size of genomes allows the generation of whole-genome methylome with relative ease. In prokaryotic microorganisms DNA methylation is playing various roles, which span from the control of cell cycle, the protection against phages (e.g. Restriction-Modification systems), and regulation of gene expression (see for examples (Sánchez-Romero and Casadesús, 2021)). Concerning cell cycle control, genome-wide DNA methylation profiles have been shown to vary in ecologically relevant contexts (e.g. bacterial differentiation, (diCenzo *et al.* , 2022)), as well as for Restriction-Modification systems strain-by-strain or population variation are documented (diCenzo *et al.* , 2022).

Consequently, the interest toward computational pipelines which can easily profile DNA methylation features in a genome-wide manner (thus allowing to compare strains and individuals across multiple conditions) is growing. Several tools have been developed for the analysis of DNA methylation profiles deriving from bisulphite sequencing and microarrays (e.g. (Müller *et al.* , 2019; Teng *et al.* , 2020; Hillary and Marioni, 2021; Aryee *et al.* , 2014; Bock *et al.* , 2005)), for a recent benchmarking see (Nunn *et al.* , 2021)). Recently, three packages have been released (Su *et al.* , 2021; Leger, 2020; De Coster *et al.* , 2020), which allow to visualize methylation profiles from SMRT or ONT sequencing data. A recent tool on GitHub has also been developed to specifically analyse DNA methylation profiles on metagenomic data (https://github.com/hoonjeseong/Meta-epigenomics). However, to the best of our knowledge, no specific pipeline has been developed for extracting DNA methylation information from sequencing data and allowing a direct quantification/comparison of the position of methylated sites with respect to genome-derived features, such as coding and noncoding sequences and report outputs which can be used in population epigenomic analyses.

Here we present MeStudio, a pipeline for SMRT sequencing methylation data integration and visualization. MeStudio combines methylation data with genome sequence and annotation to facilitate the extraction of biological information from DNA methylation profiles and to visualize the results of these analyses. We show the usage of MeStudio on a set of SMRT outputs from two strains of the bacterial species*Sinorhizobium meliloti* .

2 Design and implementation

MeStudio consists of several tools that can be run individually or as part of a pipeline and uses a naive string matching algorithm to map motif sequences to the reference genome. The required input data consist in only three files: i) a FASTA file containing the genome sequence, ii) a genomic annotation file in GFF3 format and iii) another GFF3 containing the methylated nucleotide positions. The latter is automatically generated from the output of the SMRTlink software of Pacific Biosciences DNA sequencers. As a result, MeStudio produces several files including: (i) a text file with summarized statistics concerning the methylation occurrences along the genomic features, (ii) distribution plots and, (iii) BED files containing protein annotation of the genes in which methylated motifs have been found. A workflow is provided in Figure 1.

2.1 Pre-processing

To run MeStudio, a pre-processing python script named *ms_replacR*has been implemented to produce consistent formatting on the sequence identifiers from the genomic annotation, sequencer-produced modified base calls, and the genomic sequence file. To avoid possible inconsistencies at the sequence identifiers level (the “seqid” field) between FASTA and annotation files, we have implemented a quality check in this regard. More details are provided in the MeStudio manual on GitHub.

2.2 Core-processing

The processing of the input files is handled by five executables which we refer to as “MeStudio core”. These components match the nucleotide motifs to the genomic sequence and map them to the corresponding category, which are extracted from the annotation file. Categories are defined as follows: i) protein-coding genes with accordant (sense) strand (CDS), ii) discordant (antisense) strand (nCDS), iii) regions that fall between annotated genes (true intergenic, tIG), iv) regions upstream to the reading frame of a gene, with accordant strand (US) (Figure 1B). The current implementation uses a naive matching algorithm to map motif sequences to the reference genome. During the matching stage, each replicon or chromosome gets loaded into and both strands are scanned for the presence of the motif sequences, which can hold ambiguity characters. The resulting binary files are then processed by another executable that is called for the task at hand. MeStudio core crosses methylated bases positions relative to the reference sequence start with the previously described features, producing GFF3 files that serve as input for the final analysis stage. This is a computationally expensive part of the pipeline in which multiple nested for loops and calculations are performed. Integrating one motif on a four-contigs genome (6,973,268 bp, 23,433 GANTC motif matches) took 0m27.116s on a single AMD Opteron 6380 processor (2.5GHz).

2.3 Post-processing

MeStudio implements a post-processing python script named*ms_analyzR* which takes MeStudio core output as input. In addition, to integrate comparative genomic analyses a “gene_presence_abscence.csv” file produced by Roary (Page *et al.* , 2015) can be used to define the methylation level and patterns of core and dispensable genome fractions, as well as to annotate the genes-coded proteins. *ms_analyzR* logs the total number of genes found for each category (CDS, nCDS, tIG, US). Additionally, methylation data are shown, such as i) total number of methylated sites, ii) total number of methylated genes, iii) the ID of the most methylated gene (geneID) and, iv) the product of that gene. Integrating data from Roary is functional to characterize the geneID associated with the name of the protein (as annotated by Prokka (Seemann, 2014)) as part of the core or dispensable genome. All the information is saved into a log file, together with plots accounting for the distribution of the methylations (Fig. 2A). To ensure customizability, *ms_analyzR* also includes two optional flags named "—make_chrom" and "—make_bed". The "—make_chrom" flag saves into the previously specified output directory the GFFs at "chromosome level" rather than "category level". Each GFF produced will be characterized not by category (CDS, nCDS, tIG and US) but by chromosomes (or contigs), maintaining the MeStudio core-derived contents and layout unaltered. The "—make_bed" flag produces a BED file for each feature in which is reported: i) the *chrom* column, with the name of each chromosome or contig, ii) start and iii) end of the feature, iv) the name of the geneID found in that interval, v) the number of methylations found for geneID and lastly vi) the protein product of the ID. Information contained in BED files can be readily used to plot the distribution of the methylation density for each feature, making use of the*circlize* R package (https://github.com/jokergoo/circlize) (Fig. 2B).

2.4 Tool-wide comparison

MeStudio provides a novel amount of feature-level information that is not present in other widely used genomic software packages. For instance, Bedtools (https://bedtools.readthedocs.io/en/latest/) is a well-known toolset for genomic applications through which it is possible to detect methylation features regarding CpG island, but it is not possible to extract information about CDS, nCDS, tIG and US regions as it does not provide any figure about methylated motif occurrences. Bioconductor also supplies packages that can be used for methylation analysis such as "*GenomicRanges* " (https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html) and "*motifmatchr* " (*https://bioconductor.org/packages/release/bioc/html/motifmatchr.html*). GenomicRanges allows to split the genome in predefined intervals but no information about the genomic feature is produced. The package*motifmatchr* allows to find motifs along the genome but no gene or protein annotations are included in the output. Moreover, MeStudio simply takes as input for the motifs a text file which is a more user-friendly format compared to the one required by *motifmatchr* . Table 1 provides a comparison of the features of MeStudio and possible alternative tools.

3 Case study

In order to show the performance of MeStudio, a recently published SMRT dataset was used (diCenzo *et al.* , 2022) comparing some of the methylation features of two *Sinorhizobium meliloti* strains, 1021 and FSM-MA, grown until stationary phase in minimal medium (Table 2, Figure 2B) (diCenzo *et al.* , 2022). On the SMRT assembled reads of the genomes of the two strains, MeStudio was able to identify a total of 28 motifs (Table 2). All but six motifs (namely AGAAAAT, DCTGCAGGS, RAGCWGCTY, RAGCWGCTY, RCTGCAGGS, TGGGCA) were common to both strains. The number of retrieved methylated sites ranged from a few units (especially for private motifs, those present in one strain only) to several thousands (as GANTC, which is a classical motif methylated by the CcrM DNA methylase and its involved in cell cycle regulation (Mouammine and Collier, 2018). CDS and nCDS showed similar values, as expected for methylation being present on both DNA strands. Intergenic sequences (tIG) showed the lowest number of methylated sites, while upstream sequences to a gene (UP), *bona fide* corresponding to putative promoter regions reported values generally one order of magnitude higher than tIG and in some cases differences in values between strains ranged around two-fold (e.g., CTYCCAG and GCCAGG). Finally, the presence of motifs in one strain only, may suggest the occurrence of strain-specific Restriction-Modification systems, though the small number of methylated sites may also suggest alternative hypotheses (i.e., methylation on some genomic regions only related to regulation of expression at specific loci). Demo files for input and output are available at*https://github.com/combogenomics/MeStudio.*

4 Discussion

We have reported here the description of a novel software (MeStudio) for the analysis of DNA methylation profiles obtained by single molecule real time sequencing. MeStudio has several novel and useful features compared to the few existing tools, as it provides outputs in the form of GFF and BED files which contain information on the position of methylated sites and methylated motifs, the number of methylated sites and profiles for each genomic feature and graphical outputs. The genomic features analysed include genic and intergenic regions (hence comprising putative promoters), allowing the formulation of hypotheses related to the importance of DNA methylation on regulation of gene expression and on other relevant biological phenomena. Besides being developed for prokaryotic genomes, MeStudio can handle any kind of sequence, by simply providing a suitable set of input files (Figure 1). By providing information on motif occurrence and genomic localization, MeStudio provides the basis for comparative analyses of DNA methylation profiles among strains, in terms of evolutionary studies on populations and species and epigenomic modifications during adaptation and development.

Finally, MeStudio is very user friendly given its easy installation and its possibility to be run as a pipeline, in a single command line call. We've developed the scripts in a Mac and Linux kernel environments, with the possibility in the near future to expand to Windows platforms as well.

Funding

This work was supported by MIUR, Programma Nazionale di Ricerche in Antartide 2018, (grant PNRA18_00335,*https://www.pnra.aq/*) and PRIN (Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale Escaping the ESKAPEs: integrated pipelines for new antibacterial drug, 20208LLXEJ_002) grants to MF and by the grant MICRO4Legumes, D.M.n.89267 (Italian Ministry of Agriculture) to AM. LC is supported by a PhD fellowship from MICRO4Legumes. CF is supported by a post-doctoral fellowship from the H2020 ERA-NETs SUSFOOD2 and CORE Organic Cofund, under the Joint SUSFOOD2/CORE Organic Call 2019.

Data Availability statement

The data that support the findings of this study are openly available in GitHub at https://github.com/combogenomics/MeStudio.

Conflict of Interest : none declared

References

Aryee,M.J. *et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* , **30** , 1363–1369.

Bock,C. *et al.* (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing.*Bioinformatics* , **21** , 4067–4068.

Chen,P. *et al.* (2020) Bacterial Epigenomics: Epigenetics in the Age of Population Genomics. In, Tettelin,H. and Medini,D. (eds),*The Pangenome: Diversity, Dynamics and Evolution of Genomes* . Springer, Cham (CH).

Clarke,J. *et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* ,**4** , 265–270.

De Coster,W. *et al.* (2020) Methplotlib: analysis of modified nucleotides from nanopore sequencing. *Bioinformatics* ,**36** , 3236–3238.

diCenzo,G.C. *et al.* (2022) DNA Methylation in Ensifer Species during Free-Living Growth and during Nitrogen-Fixing Symbiosis with Medicago spp. *mSystems* .

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

Flusberg,B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* ,**7** , 461–465.

Gouil,Q. and Keniry,A. (2019) Latest techniques to study DNA methylation. *Essays Biochem.* , **63** , 639–648.

Hillary,R.F. and Marioni,R.E. (2021) MethylDetectR: a software for methylation-based health profiling.

Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* , **13** , 484–492.

Leger,A. (2020) a-slide/pycoMeth.

Moler,E.R.V. *et al.* (2019) Population Epigenomics: Advancing Understanding of Phenotypic Plasticity, Acclimation, Adaptation and Diseases. In, Rajora,O.P. (ed), *Population Genomics: Concepts, Approaches and Applications* , Population Genomics. Springer International Publishing, Cham, pp. 179–260.

Mouammine,A. and Collier,J. (2018) The impact of DNA methylation in Alphaproteobacteria. *Mol. Microbiol.* , **110** , 1–10.

Müller,F. *et al.* (2019) RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* , **20** , 55.

Nunn,A. *et al.* (2021) Comprehensive benchmarking of software for mapping whole genome bisulfite data: from read alignment to DNA methylation analysis. *Brief. Bioinform.* , **22** , bbab021.

Page,A.J. *et al.* (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* , **31** , 3691–3693.

Rand,A.C. *et al.* (2017) Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* , **14** , 411–413.

Sánchez-Romero,M.A. and Casadesús,J. (2020) The bacterial epigenome.*Nat. Rev. Microbiol.* , **18** , 7–20.

Sánchez-Romero,M.A. and Casadesús,J. (2021) Waddington's Landscapes in the Bacterial World. *Front. Microbiol.* , **12** .

Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation.*Bioinformatics* , **30** , 2068–2069.

Simpson,J.T. *et al.* (2017) Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* , **14** , 407–410.

Su,S. *et al.* (2021) NanoMethViz: An R/Bioconductor package for visualizing long-read methylation data. *PLOS Comput. Biol.* ,**17** , e1009524.

Teng,C.-S. *et al.* (2020) MethGET: web-based bioinformatics software for correlating genome-wide DNA methylation and gene expression. *BMC Genomics* , **21** , 375.

Table 1 . MeStudio features compared to existing tools

Tool	**Programming language**	**Motif recognition**	**Motif matching with respect to genomic fe**
MeStudio	Python, C	yes	yes
GenomicRanges	R, C	No	No
motifmatchr	R, C++	Yes	Yes (only providing genomic ranges)
Meta-epigenomics			
Methplotlib	Python, Bash	No	No
a-slide/pycoMeth	Python, Bash	No	No
NanoMethViz	Python, Bash	No	No

Table 2. Number of methylated sites detected in S. meliloti strains FSMA-MA and 1021.

CDS, coding sequence; nCDS, coding sequence reverse strand; tIG, intergenic sequence between two genes in opposite directions; US, upstream sequence to a coding sequence. N.d., not detected.

	FSM-MA	**FSM-MA**	**FSM-MA**	**FSM-MA**	**1021**	**1021**	**1021**	**1021**
Motif	**CDS**	**nCDS**	**tIG**	**UP**	**CDS**	**nCDS**	**tIG**	**UP**
ACGGAG	50	53	7	66	64	55	8	69
AGAAAAT	N.d.	N.d.	N.d.	N.d.	6	8	1	9
BNNCGATCGV	368	397	16	364	386	450	15	408
BYCGATCG	80	119	8	125	91	110	5	114
CCCGGG	26	35	3	42	33	43	3	46
CGATCGV	405	402	19	372	409	426	16	387
CTCGAG	143	137	8	144	127	170	14	176
CTYCCAG	14	28	2	31	20	51	4	58
DCTGCAGGS	13	15	1	17	N.d.	N.d.	N.d.	N.d.
GANTC	4193	4193	1590	2719	4196	4196	1575	2731
GCCAGG	22	50	3	52	84	111	4	113
GCCGGCH	360	292	28	294	436	412	32	384
GCCGGCYD	151	143	10	151	201	189	17	197
GCRDB	3312	3234	476	1780	3714	3604	588	1929
GNCGATCGVC	97	90	4	90	113	111	2	108
RAGCWGCTY	12	16	3	21	N.d.	N.d.	N.d.	N.d.
RCCAGCC	39	64	2	68	61	70	2	70
RCGATCGGC	66	25	2	27	59	30	3	36
RCTGCAGGS	13	14	1	16	N.d.	N.d.	N.d.	N.d.
RGATCY	61	62	3	63	84	92	9	106
SCTCGAG	112	114	7	117	107	163	11	167
TCGWCGA	291	224	8	222	197	132	9	143
TGGGCA	N.d.	N.d.	N.d.	N.d.	35	32	2	36
VGCCGGCCC	11	16	2	19	20	25	3	30
VNCGATCGV	396	388	15	360	419	431	16	392
YCGATCGD	94	127	9	127	80	100	3	95
YCGGCCGRV	123	135	16	153	158	158	12	171
YCTGCAG	41	45	1	48	51	54	5	61

Figure 1 : MeStudio overview. A) Workflow. Each blue block represents input files. The green blocks indicate the scripts. The gray boxes indicate output files. B) Graphical representation of the used terminology; CDS, coding sequence; nCDS, coding sequence reverse strand; tIG, intergenic sequence between two genes in opposite directions; US, upstream sequence to a coding sequence, viz. intergenic sequence between two genes having the same orientation. See text for details.

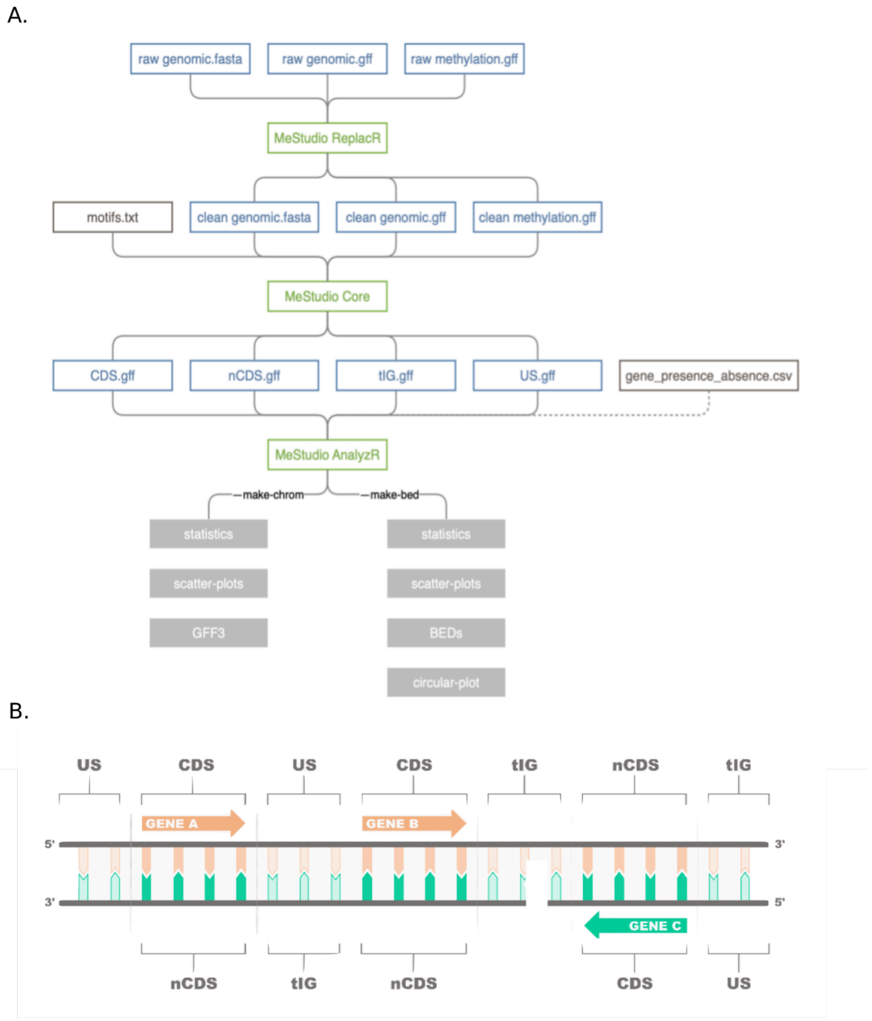
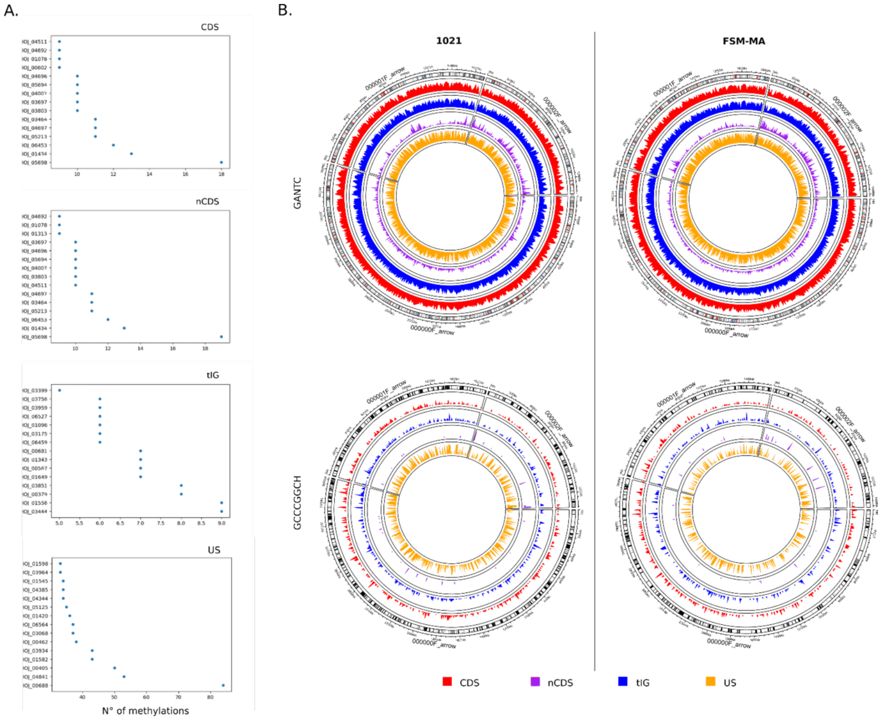
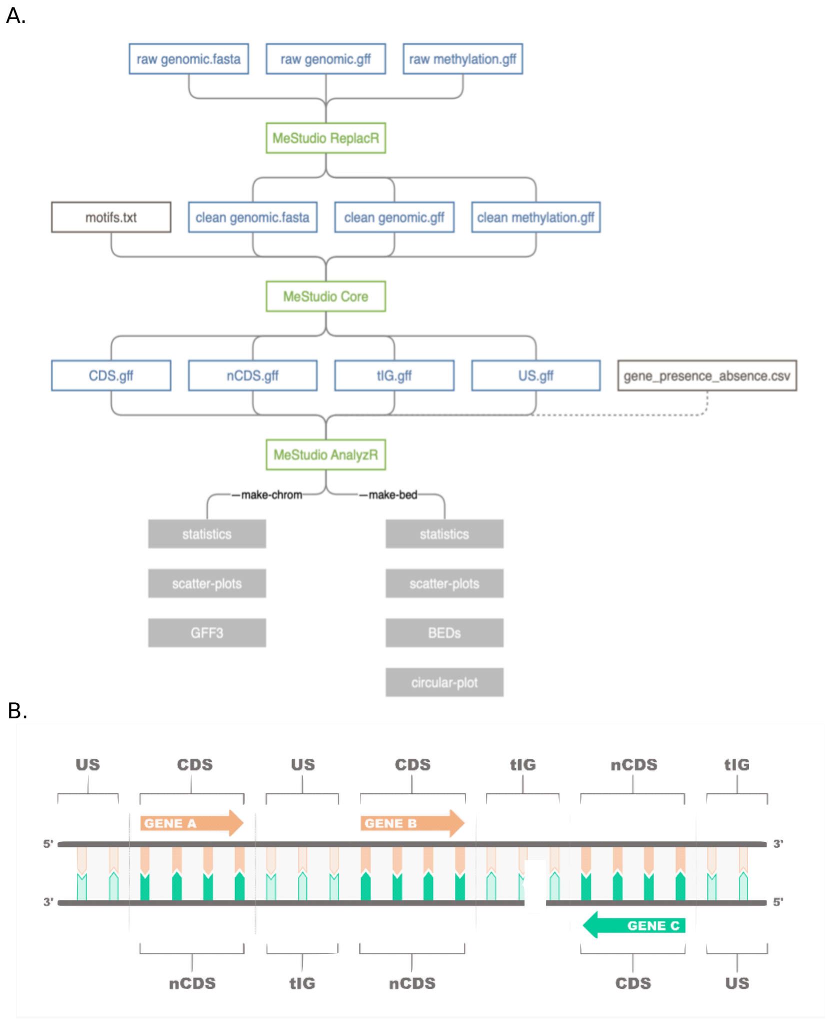
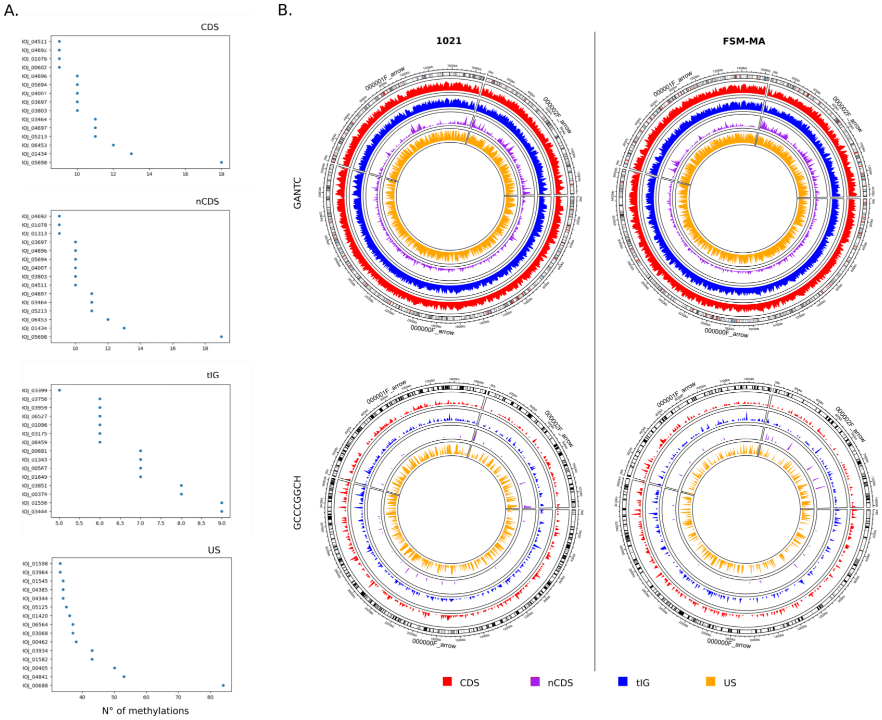


Figure 2 : Main graphical outputs of MeStudio and comparison of methylation profiles between genomes. A) Sample plot of GANTC motif in*S. meliloti* FSM-MA. Y-axis reports geneIDs, whereas X-axis reports the number of methylations found for each geneID. GeneIDs are referred to the annotation (see GitHub repository for the annotation files:*https://github.com/combogenomics/MeStudio*). B) Circular density plot of GANTC and GCCCGGCH motifs in FSM-MA and 1021 strains of *Sinorhizobium meliloti* . The outer circle represents the genome annotation of the contigs of the strain (black lines indicate the position of CDS). Each inner circle represents a different category of methylated sites, CDS (red), nCDS (blue), tIG (purple) and US (yellow). The bars of each plot indicate the values for each category.


A.
CDS
IOJ_04511
IOJ_04692
IOJ 01078
IOJ_00602
IOJ_04696
IOJ_05694
IOJ_04007
IOJ_03697
IOJ_03803
IOJ_03464
IOJ_04697
IOJ_05213
IOJ_06453
IOJ_01434
IOJ_05698
10
12
14
16
18
nCDS
IOJ_04692
IOJ_01078
IOJ 01313
IOJ_03697
IOJ_04696
IOJ_05694
IOJ_04007
IOJ_03803
IOJ_04511
IOJ_04697
IOJ 03464
IOJ_05213
IOJ_06453
IOJ 01434
IOJ_05698
10
12
14
16
18
tIG
IOJ_03399
IOJ_03756
IOJ_03959
IOJ_06527
IOJ_01096
IOJ_03175
IOJ_06459
IOJ_00681
IOJ 01343
IOJ_00567
IOJ_01649
IOJ_03851
IOJ_00379
IOJ 01556
IOJ_03444
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
US
IOJ_01598
IOJ_03964
IOJ_01545
IOJ_04385
IOJ_04344
IOJ_05125
IOJ_01420
IOJ_06564
IOJ_03068
IOJ_00462
IOJ_03934
IOJ_01582
IOJ_00405
IOJ_04841
IOJ_00688
40
50
60
70
80
N° of methylations
B.
1021
FSM-MA
GANTC
GCCCGGCH
000001F_arrow
000002F_arrow
000000F_arrow
🟥 CDS
🟪 nCDS
🟦 tIG
🟧 US


A.
raw genomic.fasta
raw genomic.gff
raw methylation.gff
MeStudio ReplacR
motifs.txt
clean genomic.fasta
clean genomic.gff
clean methylation.gff
MeStudio Core
CDS.gff
nCDS.gff
tIG.gff
US.gff
gene_presence_absence.csv
MeStudio AnalyzR
make-chrom
make-bed
statistics
scatter-plots
GFF3
statistics
scatter-plots
BEDs
circular-plot
B.
US
CDS
US
CDS
tIG
nCDS
tIG
GENE A
GENE B
5'
3'
3'
5'
GENE C
nCDS
tIG
nCDS
CDS
US


A.
CDS
IOJ_04511
IOJ_04692
IOJ_01078
IOJ_00602
IOJ_04696
IOJ_05694
IOJ_04007
IOJ_03697
IOJ_03803
IOJ_03464
IOJ_04697
IOJ_05213
IOJ_06453
IOJ_01434
IOJ_05698
10
12
14
16
18
nCDS
IOJ_04692
IOJ_01078
IOJ_01313
IOJ_03697
IOJ_04696
IOJ_05694
IOJ_04007
IOJ_03803
IOJ_04511
IOJ_04697
IOJ_03464
IOJ_05213
IOJ_06453
IOJ_01434
IOJ_05698
10
12
14
16
18
tIG
IOJ_03399
IOJ_03756
IOJ_03959
IOJ_06527
IOJ_01096
IOJ_03175
IOJ_06459
IOJ_00681
IOJ_01343
IOJ_00567
IOJ_01649
IOJ_03851
IOJ_00379
IOJ_01556
IOJ_03444
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
US
IOJ_01598
IOJ_03964
IOJ_01545
IOJ_04385
IOJ_04344
IOJ_05125
IOJ_01420
IOJ_06564
IOJ_03068
IOJ_00462
IOJ_03934
IOJ_01582
IOJ_00405
IOJ_04841
IOJ_00688
40
50
60
70
80
N° of methylations
B.
1021
FSM-MA
GANTC
GCCCGGCH
000001F_arrow
000002F_arrow
000000F_arrow
CDS
nCDS
tIG
US