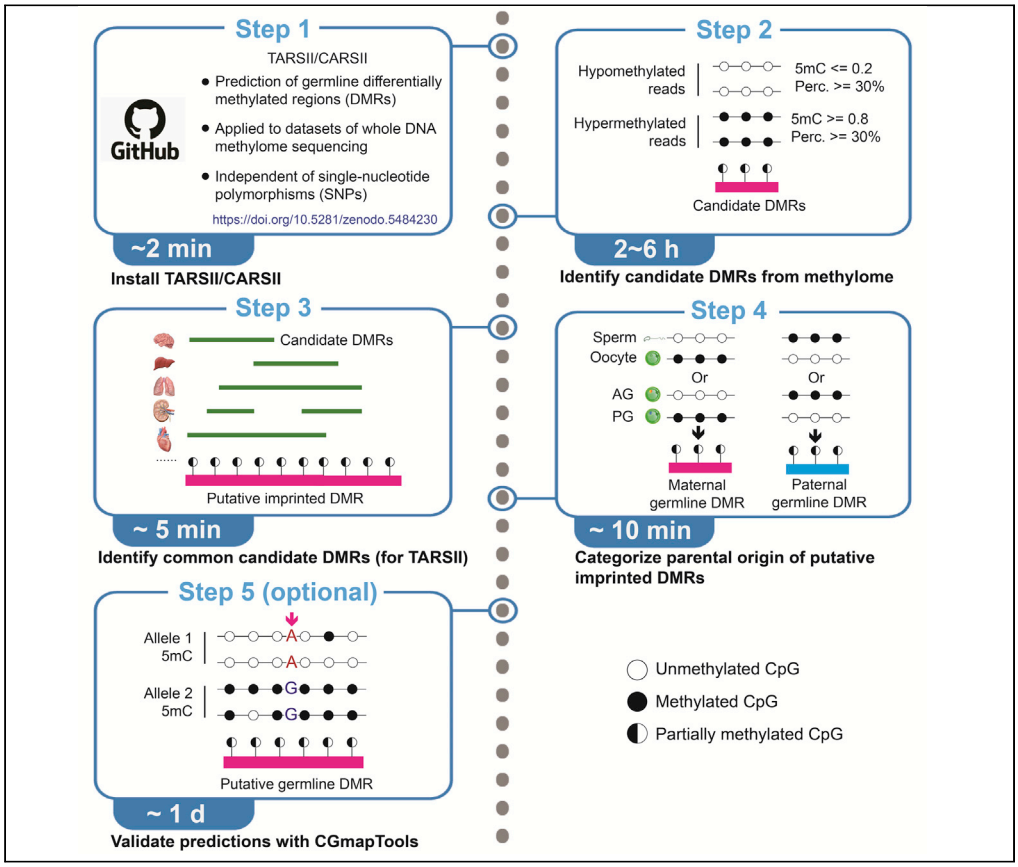
STAR Protocols

CellPress
OPEN ACCESS

Protocol

TARSII and CARSII: Two approaches for SNP-independent identification of germline differentially methylated regions in mammals



Identifying germline differentially methylated regions (DMRs) in outbred mammals remains a challenge because of difficulty in obtaining single-nucleotide polymorphisms (SNPs). To overcome this difficulty, we developed two computational approaches, TARSII and CARSII, which allow accurate prediction of germline DMRs from DNA methylomes independent of SNPs. Furthermore, we introduce an easy and quick way to validate the predicted germline DMRs with allelic DNA methylation using CGmapTools. Collectively, our strategy can greatly facilitate *de novo* identification of germline DMRs in outbred mammals.

Wenhao Zhang, Yi Zhang

yzhang@genetics.med.harvard.edu

Highlights

TARSII and CARSII predict germline DMRs from DNA methylomes independent of SNPs

Detailed protocol on how to run TARSII/CARSII and interpret the results

An easy and quick way to validate novel germline DMRs predicted by TARSII/CARSII

Zhang & Zhang, STAR Protocols 3, 101240
June 17, 2022
https://doi.org/10.1016/j.xpro.2022.101240

STAR Protocols

CellPress
OPEN ACCESS

Protocol

TARSII and CARSII: Two approaches for SNP-independent identification of germline differentially methylated regions in mammals

Wenhao Zhang[1,2,3,6] and Yi Zhang[1,2,3,4,5,7,*]

[1]Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA 02115, USA

[2]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA 02115, USA

[3]Division of Hematology/Oncology, Department of Pediatrics, Boston Children's Hospital, Boston, MA 02115, USA

[4]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

[5]Harvard Stem Cell Institute, WAB-149G, 200 Longwood Avenue, Boston, MA 02115, USA

[6]Technical contact

[7]Lead contact

*Correspondence: yzhang@genetics.med.harvard.edu

https://doi.org/10.1016/j.xpro.2022.101240

SUMMARY

Identifying germline differentially methylated regions (DMRs) in outbred mammals remains a challenge because of difficulty in obtaining single-nucleotide polymorphisms (SNPs). To overcome this difficulty, we developed two computational approaches, TARSII and CARSII, which allow accurate prediction of germline DMRs from DNA methylomes independent of SNPs. Furthermore, we introduce an easy and quick way to validate the predicted germline DMRs with allelic DNA methylation using CGmapTools. Collectively, our strategy can greatly facilitate *de novo* identification of germline DMRs in outbred mammals.

For complete details on the use and execution of this protocol, please refer to Chu et al. (2021).

BEFORE YOU BEGIN

Genomic imprinting is the preferential expression of one of the parental alleles, with silencing of the other allele driven by epigenetics (Barlow and Bartolomei, 2014). Although both allelic DNA methylation and H3K27me3 mediate genomic imprinting, allelic DNA methylation plays a major role for imprinting regulation in somatic tissues (Barlow and Bartolomei, 2014; Chen and Zhang, 2020; Inoue et al., 2017). Although allelic DNA methylation from parental alleles can be calculated using strain-specific single-nucleotide polymorphisms (SNPs) in inbred species, such strategy is difficult to be applied to outbred mammals such as human and monkey. Currently, *de novo* identification of genome-wide allelic differentially methylated regions (DMRs) in human requires integrating hundreds of methylomes from different individuals based on SNP analysis (Zink et al., 2018) or studying special samples such as those with uniparental disomy (Joshi et al., 2016). These methods are both resource and time consuming and thus cannot be widely used in outbred mammals.

To overcome this difficulty, we have developed two computational approaches TARSII (tissue-associated, reads-based, SNP-free approach for identifying imprint-DMRs) and CARSII (CpG-island-associated, reads-based, SNP-free approach for identifying imprint-DMRs), which allow prediction of germline DMRs from whole DNA methylomes generated by next-generation sequencing without SNPs. TARSII is capable of accurately identifying genome-wide germline DMRs using as few as six DNA methylomes from different somatic tissues, which can be derived from either the same individual or different individuals. CARSII is designed to predict germline DMRs from a single DNA

methylome. DNA methylomes of sperm and oocyte, or that of uniparental early embryos are also required for categorizing the parental origin of the predicted imprinted DMRs in SNP-independent manner.

Although our approaches can predict germline DMRs with a relative high accuracy, validation by allelic methylation analyses is recommended, especially for predicting novel germline DMRs. To this end, we introduce an easy and quick method based on the CGmapTools (Guo et al., 2018). CGmapTools identifies SNPs directly from the DNA methylome and can calculate allelic DNA methylation associated with individual SNP (Guo et al., 2018). By first predicting germline DMRs in genome-wide level using TARSII/CARSII and then validating the predicted germline DMRs using a few SNPs inside, novel germline DMRs can be easily and efficiently identified in outbred mammals with limited samples.

Prepare your DNA methylome datasets

Timing: 1–2 weeks

Note: Our instruction to process the DNA methylome datasets for TARSII and CARSII is based on Linux system. The timing is calculated based on a computer with 64 Gb memory and 4 cores, which is also the minimal requirement.

To properly execute TARSII and CARSII, the users will need to first prepare DNA methylomes and generate the required files as instructed below. We recommend using DNA methylomes from at least six different somatic tissues derived from all three germ layers. In addition, DNA methylomes of sperm/oocyte or uniparental early embryos are also required to categorize the parental origin of those predicted imprinted DMRs. Currently, only reads aligned with Bismark and deduplicated with Picard tools are compatible to TARSII/CARSII.

On the other hand, some software (TARSII/CARSII, Bismark, Picard tools, CGmapTools, Trim galore, see also key resources table) are required to be installed in computer before processing. To install the software, please refer to the tutorials by clicking the related links under the IDENTIFIER column in key resources table.

The users can process the raw reads of DNA methylomes for TARSII/CARSII following the steps below:

1. Trim sequencing adaptors of the FASTQ file using Trim Galore.
   a. For data generated with traditional whole genome bisulfite sequencing (WGBS) method:

```
# Take 5mC.fastq as an example of the input fastq file

# if standard sequencing adaptor of Illumina is used:

Trim_galore 5mC.fastq -illumina

# if custom adaptor is used:

Trim_galore 5mC.fastq -adaptor custom_adaptor
```

   b. For data generated with post-bisulfite adaptor tagging (PBAT):

   Note: Removal of the first and last 9 base pairs in PBAT with "–clip_R1 9" and "–three_prime_clip_R1 9" parameters is recommended to reduce the influence of the random sequences added during library preparation.

```
# Take 5mC.fastq as an example of the input fastq file
Trim_galore 5mC.fastq –dont_gzip –clip_R1 9 –three_prime_clip_R1 9
```

2. Aligning reads with Bismark following the command:

```
# Input fastq file should be trimmed before processing
bismark –fastq -L 30 -N 1 –non_directional genome_build_folder
5mC_trimmed.fastq
```

3. Remove the PCR duplicates with Picard tools and generate sam file following the commands:

```
java -Xmx32g -jar picard.jar SortSam INPUT=5mC_bismark_bt2.bam
OUTPUT=picard_sorted.bam SORT_ORDER=coordinate CREATE_INDEX=true
# Then
java -Xmx32g -jar picard.jar MarkDuplicates INPUT=picard_sorted.bam
METRICS_FILE=metrics.txt REMOVE_DUPLICATES=true ASSUME_SORTED=true
TMP_DIR=tmp_dir VALIDATION_STRINGENCY=SILENT CREATE_INDEX=true
# The deduplicated reads generated above will be stored in the file
named picard_deduplicated.bam, which can be converted to sam file with
the following command:
samtools view picard_deduplicated.bam > picard_deduplicated.sam
```

4. Generate wig file of DNA methylation at base resolution with Bismark following the commands:

```
# Extract methylated cytosines from input bam file
bismark_methylation_extractor –no_overlap –report –bedGraph –cutoff
3 picard_deduplicated.bam
# Calculate DNA methylation level
coverage2cytosine –merge_CpG –genome_folder genome_build_folder -o
output picard_deduplicated.bismark.cov.gz
# Generate wig file with DNA methylation level at base resolution
awk '{printf ''%s\t%d\t%.2f\n'', $1, $2, $4/100}'
output.CpG_report.merge_CpG_evidence.cov > 5mC.wig
```

Note: For each CpG site, we require its methylated state detected by at least three times, which can be fulfilled by adding parameter "–cutoff 3" in "bismark_methylation_extractor" command. We also require CpGs of both forward and reverse strands to be merged, which can be fulfilled by adding parameter "–merge_CpG" in "coverage2cytosine" command.

A

Input sam file processed from Bismark and MarkDuplicates

```
7001405:1143:HWJYYBCX2:2:2211:4000:26071_1:N:0:GTGGCC   0
chr3    5       32      81M     *       0       0       TTTTAGTA
GAGATAGGGTTTTATTATGTTGGTTAGGTTGTTTTGGAATTTTTGATTTTGTGATTTAATTATT
TTGGTTTTT        IIIHFIIHHGHFIIIIIIIIIHIIIIIIIIIIIHIIIIIIIIIIIHI
IIIIIIIIIIIIIIIIIIHHHIIIIIIIIHHI        MD:Z:12C7C1C0C7C0C3C2C1C
5C1C0C3C0C1C5C0C2C0C1C0C4C0C1C0C0       PG:Z:MarkDuplicates
XG:Z:CT NM:i:25 XM:Z:............x.......h.hh.......hx...x..h.x.
....h.hx...hh.z.....hh..hh.hh....hh.hh  XR:Z:GA
```

B

Input wig file showing DNA methylation levels at base resolution

Chromatin	CG Position	Methylation
chr1	12363	0.50
chr1	12415	0.33
chr1	13198	0.00
chr1	13272	0.25
chr1	13572	0.08
chr1	14447	0.29

Figure 1. Format example of the input files required by TARSII and CARSII

(A) Format example of the input sam file prossed from Bismark and MarkDuplicates.

(B) Format example of the input wig file containing DNA methylation levels at base resolution.

The users can refer to more detailed instruction in Bismark by clicking on this link (https://github.com/FelixKrueger/Bismark/tree/master/Docs) or generate the wig file using customized pipelines. The file for DNA methylation level is tab separated with columns of chromatin, CG base position (merge both forward and reverse strands) and DNA methylation levels (ranges from 0 to 1) (Figure 1).

⚠ CRITICAL: Notably, traditional WGBS method requires to first break genomic DNA into fragments, ligate adaptors to the fragments, and then perform bisulfite treatment to the ligated fragments. This strategy may cause loss of DNA methylation in part of the reads of the read 2 file in pair-end sequencing. Thus, if the WGBS library is constructed using the way described above, we suggest to either only use the read 1 file or remove the reads in the read 2 file with sam Flag marked as 177 or 129 (two reads mapped to the same DNA strand). If the genomic DNA is first processed by bisulfite treatment and then ligated to adaptors as used in PBAT, both read 1 and read 2 files in pair-end sequencing can be included. In that case, the two fastq files in pair-end sequencing can be merged and processed as a single file.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Public DNA methylomes of mouse somatic tissues	(Hon et al., 2013)	GEO: GSE42836
Public DNA methylomes of mouse gemmates	(Wang et al., 2014)	GEO: GSE56697
Public DNA methylomes of human somatic tissues	(Court et al., 2014)	GEO: GSE52578
Public DNA methylomes of human somatic tissues	NCBI's Gene Expression Omnibus	GEO: GSE16256
Public DNA methylomes of human uniparental early embryos	(Leng et al., 2019)	GEO: GSE133856
Public DNA methylomes of monkey somatic tissues and uniparental early embryos	(Chu et al., 2021)	GEO: GSE159347
Script in this protocol	Mendeley Data	Mendeley Data: https://doi.org/10.17632/747r4k4mnz.1

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
TARSII/CARSII	(Chu et al., 2021)	https://doi.org/10.5281/zenodo.5484230
Bismark	(Krueger and Andrews, 2011)	https://github.com/FelixKrueger/Bismark
Picard Tools	Broad Institute	http://broadinstitute.github.io/picard/
CGmapTools	(Guo et al., 2018)	https://cgmaptools.github.io/
Trim Galore	Felix Krueger in Babraham Institute	https://doi.org/10.5281/zenodo.5127899

MATERIALS AND EQUIPMENT

Input datasets

To get started, the users can use the public methylomes of somatic tissues and gemmates/uniparental embryos from mouse, human or monkey. These datasets were used and can be found in our previous study (Chu et al., 2021). For convenience, the accession numbers of these datasets are also listed in the key resources table.

Here, we use the human DNA methylomes (brain, muscle, aorta, lung, liver, intestine, androgenetic 8-cell embryo, parthenogenetic 8-cell embryo) as a working example.

Input files:

- Sam file: DNA methylomes of somatic tissues should be processed into sam file by alignment with Bismark and removing duplicates with MarkDuplicate in Picard Tools as instructed above (Figure 1A).
- Wig file: DNA methylomes of somatic tissues, sperm/oocyte or uniparental early embryos should be processed into wig file showing DNA methylation levels at base resolution as instructed above (Figure 1B)

STEP-BY-STEP METHOD DETAILS

Install TARSII and CARSII

Timing: 2 min

Note: The timing calculated for each step of TARSII and CARSII is based on a computer with 8 Gb memory and 1 core. A minimal of 8 Gb memory and 1 core is required. TARSII and CARSII are functional on Linux and Mac OSX systems.

1. The TARSII/CARSII packages are available through the following public repository:

https://doi.org/10.5281/zenodo.5484230

To activate TARSII and CARSII, you may follow the command lines below:

```
cd TARSII/CARSII_folder

chmod u+x *

cd ./bin

chmod u+x *
```

Predict germline DMRs using TARSII

Considering germline DMRs are composed of both hypomethylated and hypermethylated alleles, those regions should have partial DNA methylation and should enrich for both hypomethylated and hypermethylated reads (Figure 2A). Because DMRs are believed to be generally conserved across different tissues (Babak et al., 2015), identifying common candidate DMRs across different tissues can help reduce FDR. These are the basis for imprinting prediction by TARSII.

Identify candidate DMRs from individual DNA methylome by TARSII

Timing: 2 h

In this step, candidate DMRs in each tissue are selected out by two analyses. First, genomic regions containing consecutive CpG sites (>= 10) with 5mC level ranging from 0.3 to 0.7 are selected out as partially methylated domains (PMDs) (Figure 2B). Then, the PMDs enriched for both hypomethylated reads (5mC <= 0.2, reads percentage >= 30%) and hypermethylated reads (5mC >= 0.8, reads percentage >= 30%) are identified as candidate DMRs (Figure 2B).

2. Candidate DMRs of individual somatic tissue methylome are identified following the command:

```
# Take cortex methylome data as an example

TARSII_step1_DMR_identify.sh -x cortex_5mC.wig -s

cortex_picard_deduplicated.sam -o cortex

# Run this script for DNA methylome of each somatic tissue
```

Note: The input files and output file name are mandatory; other parameters have been optimized but can be adjusted according to user-specific requests.

Parameters should be provided:

-x wig file presenting DNA methylation levels at base resolution

-s sorted sam file with duplicates removed

-o output file name

Parameters available to be adjusted through [options]:

-n minimal CpG number required to be included in a PMD. Default: 10 (>= 1)

-m minimal methylation level for the CpG sites in a PMD. Default: 0.3 (ranges from 0 to 1)

-M maximal methylation level for the CpG sites in a PMD. Default: 0.7 (ranges from 0 to 1)

-r minimal CpG number required in a single read. Default: 3 (>= 1)

-l minimal number of reads required to be aligned to a PMD. Default: 30 (>=1)

-b bin number to categorize methylation levels of the reads in a PMD. Default: 5 (>= 2)

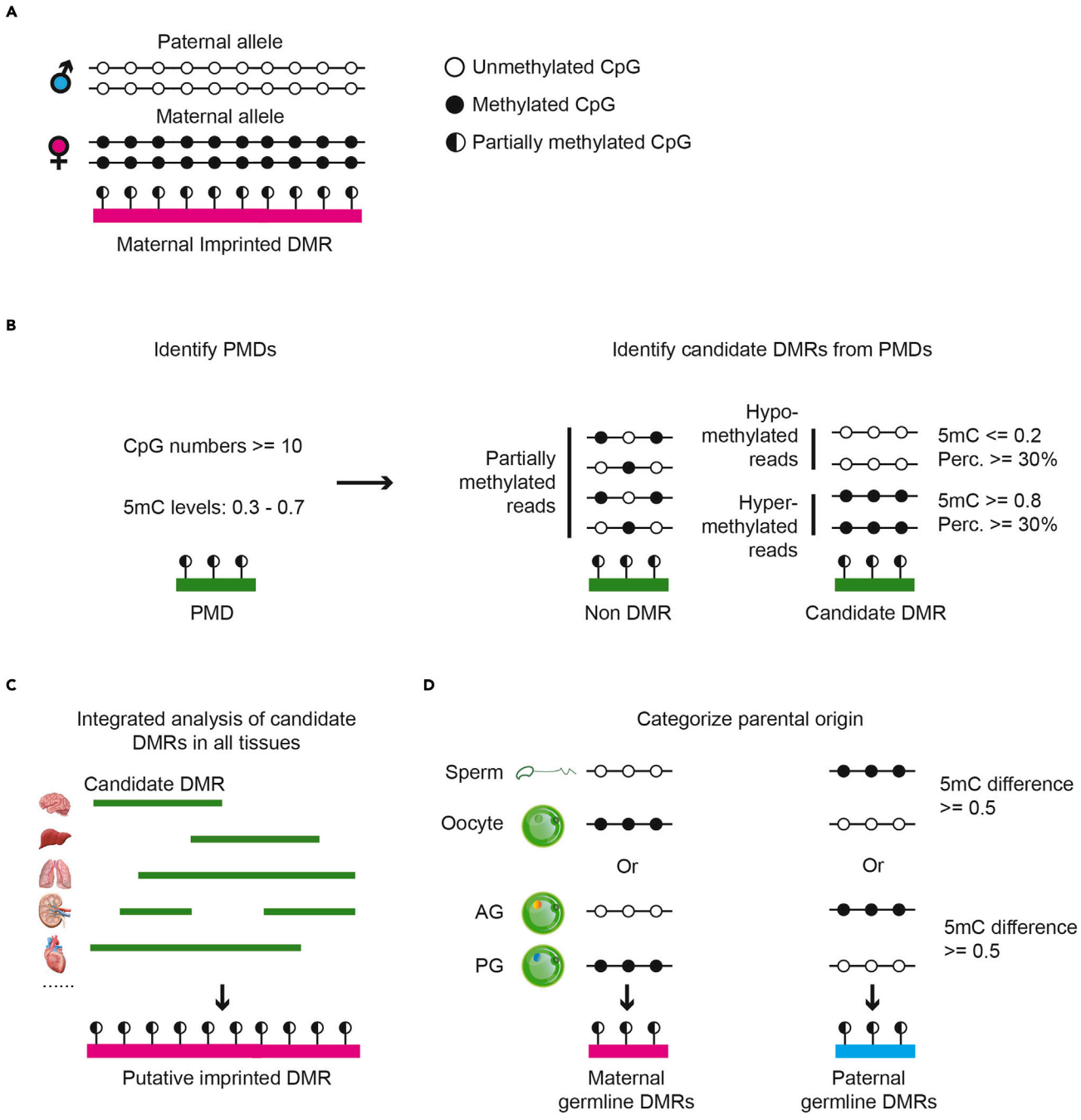


Figure 2. Model for prediction of putative germline DMRs by TARSII

(A) Schematic model showing allelic DNA methylation for a typical maternal imprinted DMR.

(B) Schematic models showing the strategies for identifying PMDs from the genome and for identifying candidate DMRs from the PMDs in step 2.

(C) Schematic model showing the strategy for integrated analysis of candidate DMRs in all tissues in step 4.

(D) Schematic model showing the strategy for categorizing the parental origin of the predicted imprinted DMRs in step 7.

Note: -b 5 means to categorize the reads into 5 groups with methylation levels range from 0.0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1.0. Option -c/-C is applied to the first/last bin

-c minimal percentages of hypomethylated reads versus total reads for a candidate DMR. Default: 0.3 (30%) (ranges from 0 to 1)

-C minimal percentages of hypermethylated reads versus total reads for a candidate DMR. Default: 0.3 (30%) (ranges from 0 to 1)

3. Following completion of step 2, users will get 3 output files in working directory that include (troubleshooting problem 1):
   a. A bed file containing all PMDs identified from the input DNA methylome.
   b. A tab separated file containing all PMDs with additional information as percentage of hypomethylated reads, percentage of hypermethylated reads and total reads number in each bin.
   c. A bed file containing all candidate DMRs identified from the input DNA methylome for processing in the next step.

Identify common candidate DMRs from different somatic tissues

Timing: 5 min

In this step, the overlapping candidate DMRs from all tissues are merged as one (Figure 2C). Then, if the candidate DMRs from at least 5 different tissues show overlap with the merged candidate DMRs, those candidate DMRs are defined as putative imprinted DMRs identified by TARSII (Figure 2C).

4. The putative imprinted DMRs are identified through integrated analysis of candidate DMRs in different somatic tissues following the command:

```
TARSII_step2_DMR_integration.sh -f 'cerebellum_DMR_candidate.bed
cortex_DMR_candidate.bed heart_DMR_candidate.bed
intestine_DMR_candidate.bed kidney_DMR_candidate.bed
liver_DMR_candidate.bed' -o human
```

Note: Users should include all the bed files of candidate DMRs and an output file name to run this script. The minimal number of tissues required for a putative imprinted DMR to be identified could be adjusted with the parameter -n. However, users should be cautious that reducing the cutoff will increase false discovery rate, while increasing the cutoff may increase the false negative rate.

Parameters should be provided:

-f all bed files containing candidate DMRs identified from step 1

Note: File names should be located within ' ' symbol

-o output file name

Parameters available to be adjusted through [options]:

-n minimal number of tissues for a putative imprinted DMR to be commonly identified in (>= 1)

5. Following completion of step 4, users will get 1 bed file presenting the putative imprinted DMRs in working directory.

Categorize parental origin of the putative imprinted DMRs predicted by TARSII

Timing: 10 min

Since the prediction of TARSII is independent of SNPs, the parental origin of these imprinted DMRs cannot be directly inferred from somatic tissue methylomes by TARSII. Nevertheless, considering the allelic DNA methylation of germline DMRs is inherited from gametes, the differentially methylated regions between sperm and oocyte, as well as between androgenetic and parthenogenetic early embryos, can help infer the parental origin of the predicted imprinted DMRs without SNP information.

In this step, paternal and maternal DMRs in sperm/oocyte or uniparental early embryos are identified. Based on those parental DMRs, TARSII predicted putative imprinted DMRs can be categorized into maternal germline DMRs, paternal germline DMRs and somatic DMRs (Figure 2D).

6. Wig files showing DNA methylation levels at base resolution from sperm/oocyte or uniparental early embryos are required, as shown in Figure 1B. To prepare for these files, please refer to "before you begin" and our previous study (Chu et al., 2021).
7. Parental origin of the putative imprinted DMRs predicted by TARSII can be categorized following the command:

```
TARSII_step3_germline_DMR.sh -p androgenetic_5mC.wig -m
parthenogenetic_5mC.wig -b human_putative_imprinted_DMR.bed -o human
```

Note: The input files and output file name are mandatory; other parameters have been optimized but can be adjusted according to user-specific requests.

Parameters should be provided:

-p wig file presenting DNA methylation levels at base resolution of sperm/androgenetic embryos

-m wig file presenting DNA methylation levels at base resolution of oocyte/parthenogenetic embryos

-b bed file of the imprinted DMRs predicted by TARSII (generated from step 4)

-o output file name

Parameters available to be adjusted through [options]:

-d minimal cutoff to define a differentially methylated CpG site for a DMR in gemmates. Default: 0.5 (ranges from 0 to 1)

-n minimal CpG number required to be included in a DMR in gemmates. Default: 10. (>= 1)

-c maximal paternal methylation level in a maternal DMR in gemmates. Default: 0.15 (ranges from 0 to 1)

-C maximal maternal methylation level in a paternal DMR in gemmates. Default: 0.30 (ranges from 0 to 1)

8. Following completion of step 7, users will get 5 files in working directory. File 3, 4, 5 (below) are the final results generated by TARSII (troubleshooting problems 2 and 3).
   a. A bed file contains paternal DMRs in gemmates with average methylation levels of maternal and paternal alleles.
   b. A bed file contains maternal DMRs in gemmates with average methylation levels of paternal and maternal alleles
   c. A bed file contains maternal germline DMRs predicted by TARSII.
   d. A bed file contains paternal germline DMRs predicted by TARSII.
   e. A bed file contains somatic DMRs predicted by TARSII.

Predict germline DMRs using CARSII

TARSII predicts the germline DMRs through integrated analyses of DNA methylomes from different somatic tissues. However, in some case, identification of tissue-specific imprinting is required. To help predict germline DMRs in a single tissue independent of SNPs, we developed another computational tool, CARSII.

Different from the genome-wide identification of germline DMRs in TARSII, only CpG islands (genomic regions with high CpG density) are included in the analysis of CARSII. This is mainly because: 1) DNA methylation in CpG islands is generally under rigid regulation of multiple transcription factors and epigenetic regulators (Deaton and Bird, 2011). Thus, the DNA methylation in CpG islands is more stable across different cells compared to that of a random region in genome; 2) the majority of germline DMRs in mammals, such as mouse and human, overlap with GpG islands. In contrast, CpG islands only occupy a small portion of the genome (Chu et al., 2021). Thus, the chances for a germline DMR to be identified from a CpG island are much higher than that from a random region in the genome.

Notably, the definition of CpG islands may vary according to different standards used. Nevertheless, we recommend using the CpG islands defined in the UCSC Genome Browser database, which could be found in the following link:

https://hgdownload.soe.ucsc.edu/downloads.html

Identify candidate differentially methylated CpG-islands (DMCs) from a single DNA methylome by CARSII

Timing: 6 h

The biological basis of CARSII is similar to TARSII, which assumes germline DMRs are enriched for both hypomethylated and hypermethylated alleles (Figure 2A). However, due to heterogeneity of the cells in a tissue, methylation inconsistency in part of the CpG islands and experimental variations caused by batch effects, the false discovery rate (FDR) of CARSII is relatively higher than that of TARSII. Thus, to help reduce the FDR in CARSII, we designed additional steps to remove germline DMRs that may be resulted from random effect or methylation inconsistency.

In this step, CpG islands that enriched for both hypomethylated reads (5mC <= 0.2, reads percentage >= 30%) and hypermethylated reads (5mC >= 0.8, reads percentage >= 30%) are first selected as candidate DMCs (Figure 3A). Then, random test is performed to calculate the FDR for each candidate DMC and those with FDR < 0.05 were removed (Figure 3B). Finally, the methylation consistency in each candidate DMC is determined and candidate DMCs with inconsistent DNA methylation along the CpG island are removed (Figure 3C).

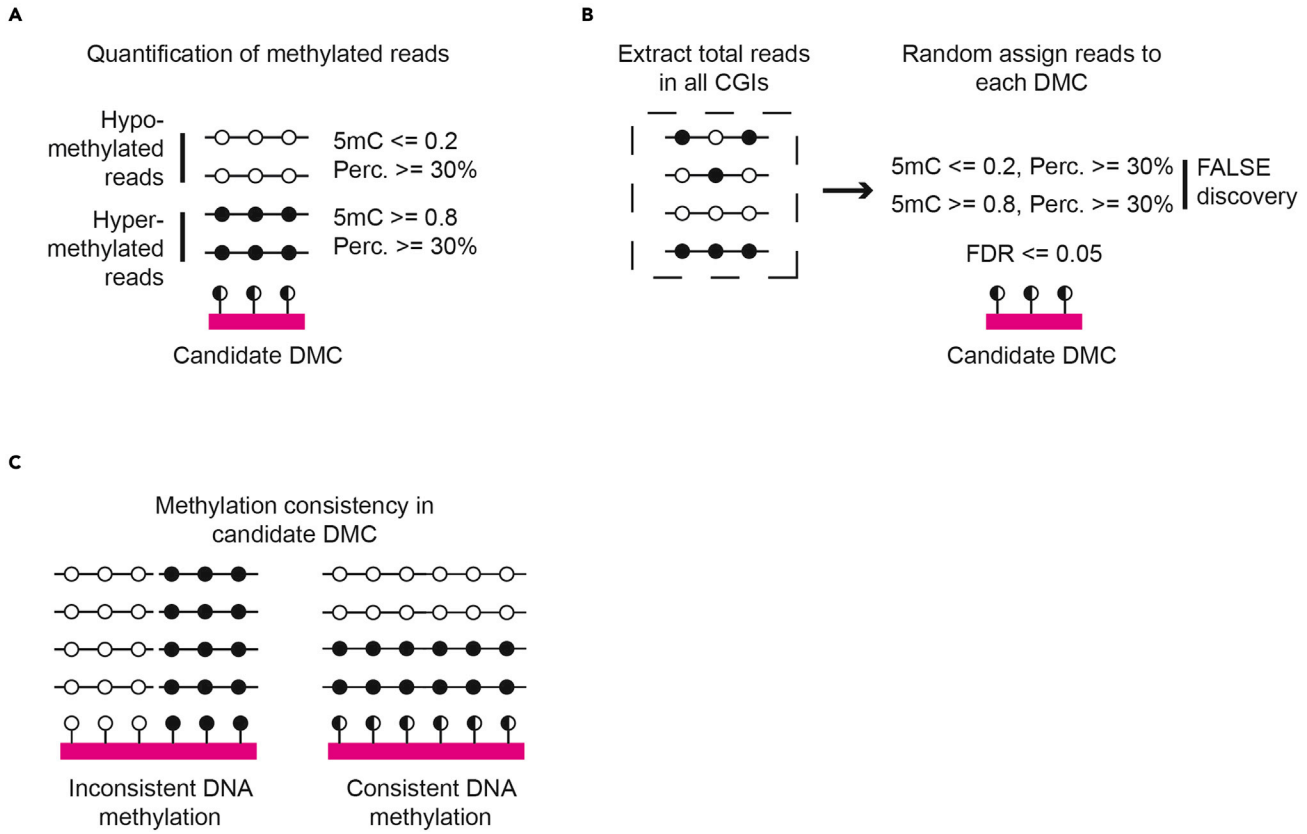


Figure 3. Model for prediction of putative germline DMCs by CARSII

(A–C) Schematic models showing the strategies for quantifying methylated reads (A), random testing reads distribution (B) and analyzing methylation consistency (C) in candidate DMCs.

9. Candidate DMCs from DNA methylome of a certain somatic tissue are identified by CARSII following the command:

```
# Take cortex methylome data as an example
CARSII_step1_DMC_identify.sh -g human_CpG_island.bed -x cortex_5mC.wig
-s cortex_picard_deduplicated.sam -o cortex
```

Note: The input files and output file names are mandatory; other parameters have been optimized but can be adjusted according to user-specific requests:

Parameters should be provided:

-g bed file of all the CpG islands

-x wig file presenting DNA methylation levels at base resolution

-s sorted sam file with duplicates removed

-o output file name

Parameters available to be adjusted through [options]:

-r minimal CpG number required in a single read. Default: 3 (>= 1)

-l minimal number of reads required to be aligned to a CpG island. Default: 20 (>= 1)

-b bin number to categorize methylation levels of the reads in a CpG island. Default: 5 (>= 2)

Note: -b 5 means to categorize the reads into 5 groups with methylation levels range from 0.0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1.0. Option -c/-C is applied to the first/last bin

-c minimal percentages of hypomethylated reads versus total reads for a candidate DMC. Default: 0.3 (30%) (ranges from 0 to 1)

-C minimal percentages of hypermethylated reads versus total reads for a candidate DMC. Default: 0.3 (30%) (ranges from 0 to 1)

-p maximal false discovery rate for a candidate DMC. Default: 0.05 (<= 0.05)

-t test times for calculation of false discovery rate. Default: 10000 (>=1)

-d maximal methylation differences allowed within a candidate DMC. Default: 0.2 (ranges from 0 to 1)

⚠ CRITICAL: We do not recommend using CARSII to directly predict non-germline/somatic DMRs because no allelic DNA methylation in gametes could be used to confirm the imprinting status for a somatic DMR.

10. Following completion of step 9, the user will get 2 files in working directory (troubleshooting problem 1):
    a. A tab separated file containing all CpG islands with the percentage of hypomethylated reads, percentage of hypermethylated reads, total reads number in a CpG island and reads number in each bin.
    b. A bed file containing predicted candidate DMCs by CARSII. The candidate DMCs in this file are viewed as the putative imprinted DMCs predicated by CARSII.

Categorize parental origin of the putative imprinted DMCs predicted by CARSII

Timing: 10 min

Similar to TARSII, in this step, paternal and maternal-specific methylated CpG-islands in gemmates are identified by analyzing DNA methylomes in sperm/oocyte or uniparental early embryos (Figure 2D). Then, based on those parental-specific methylated CpG-islands, putative imprinted DMCs are categorized into maternal germline DMCs, paternal germline DMCs and somatic DMCs (Figure 2D).

11. Parental origin of imprinted DMCs predicted by CARSII can be identified following the command:

```
# Take cortex methylome data as an example

CARSII_step2_germline_DMR.sh -p androgenetic_5mC.wig -m

parthenogenetic_5mC.wig -b cortex_putative_imprinted_DMC.bed -o cortex
```

Note: The input files and output file names are mandatory; other parameters have been optimized but can be adjusted according to user-specific requests.

Parameters should be provided:

-p wig file presenting DNA methylation levels at base resolution of sperm/androgenetic embryos

-m wig file presenting DNA methylation levels at base resolution of oocyte/parthenogenetic embryos

-b bed file of putative imprinted DMCs generated from step 1 script

-o output file name

Parameters available to be adjusted through [options]:

-d minimal cutoff to define a differentially methylated CpG site for a DMC in gemmates. Default: 0.5 (ranges from 0 to 1)

-c maximal paternal methylation in a maternal DMC in gemmates. Default: 0.15 (ranges from 0 to 1)

-C maximal maternal methylation in a paternal DMC in gemmates. Default: 0.30 (ranges from 0 to 1)

12. Following completion of step 11, the user will get 3 files in working directory (troubleshooting problems 2 and 4):
    a. A bed file containing maternal germline DMCs predicted by CARSII
    b. A bed file containing paternal germline DMCs predicted by CARSII.
    c. A bed file containing somatic DMCs predicted by CARSII.

Note: As noted above, we do not recommend applying CARSII to predict somatic DMCs. Nevertheless, if the user does use this approach to predict somatic DMCs, validation by allelic methylation analysis is suggested before moving forward (see the following part).

Analysis of allelic DNA methylation for the putative imprinted DMRs with CGmapTools

Although the putative germline DMRs predicted by TARSII and CARSII is relatively accurate (Chu et al., 2021), certain level of false discovery rate is unavoidable without the information of the allelic DNA methylation. Therefore, we introduce CGmapTools (Guo et al., 2018) to help easily and quickly validate the imprinted DMRs predicted by TARSII/CARSII. CGmapTools is capable of identifying SNPs directly from DNA methylomes and calculating allelic methylation levels associated with those SNPs.

By combining TARSII/CARSII and CGmapTools, putative imprinted DMRs in the whole genome can be identified first using TARSII/CARSII independent of SNPs. Then, allelic methylation of the putative imprinted DMRs can be calculated by CGmapTools with only a few SNPs located within those DMRs. In this way, instead of collecting large number of DNA methylomes from many different individuals and SNPs from their parents' genomes, only a few DNA methylomes are sufficient for accurate *de novo* identification of imprinted regions in outbred mammals.

Here, we only provide an integrated analysis of CGmaptools and TARSII/CARSII. For a complete and detailed instructions on CGmapTools, please refer to the published study (Guo et al., 2018) and the link below:

https://cgmaptools.github.io/

Identify SNPs from DNA methylome using CGmapTools

Timing: 1 day

In this step, SNPs in certain DNA methylome are extracted using CGmapTools (Figure 4A). In TARSII, we can extract SNPs from every DNA methylome that used for analysis.

13. SNPs from DNA methylome can be extracted following the commands:

```
# Take cortex methylome data as an example
cgmaptools convert bam2cpmap -b cortex_picard_deduplicated.bam -g
human_genome.fa -o cortex
# Then:
cgmaptools snv -i cortex.ATCGmap.gz -m bayes -v cortex_SNP.vcf -o
cortex_SNP.snv -bayes-dynamicP
```

14. Extracting SNPs located within the putative imprinted DMRs predicted by TARSII/CARSII using the perl script (Data S1) and command:

```
# Take cortex methylome data as an example
Extract_SNPs_from_DMRs.pl cortex_SNP.vcf cortex_SNP.snv
human_putative_imprinted_DMR.bed
```

This script generates a selected vcf file containing SNPs located within the putative imprinted DMRs as listed in the input bed file.

Note: CGmapTools perform allelic analysis but cannot distinguish parental origin. To infer the parental origin of non-germline/somatic DMRs, SNPs identified from parental genomes are still needed.

Calculate allelic DNA methylation using CGmapTools

Timing: 5 min

15. Allelic DNA methylation is calculated by CGmapTools following the command:

```
# Take cortex methylome data as an example
cgmaptools asm -r human_genome.fa -b cortex_picard_deduplicated.bam -l
cortex_SNP_DMRs.vcf > cortex_SNP_DMRs.asm
```

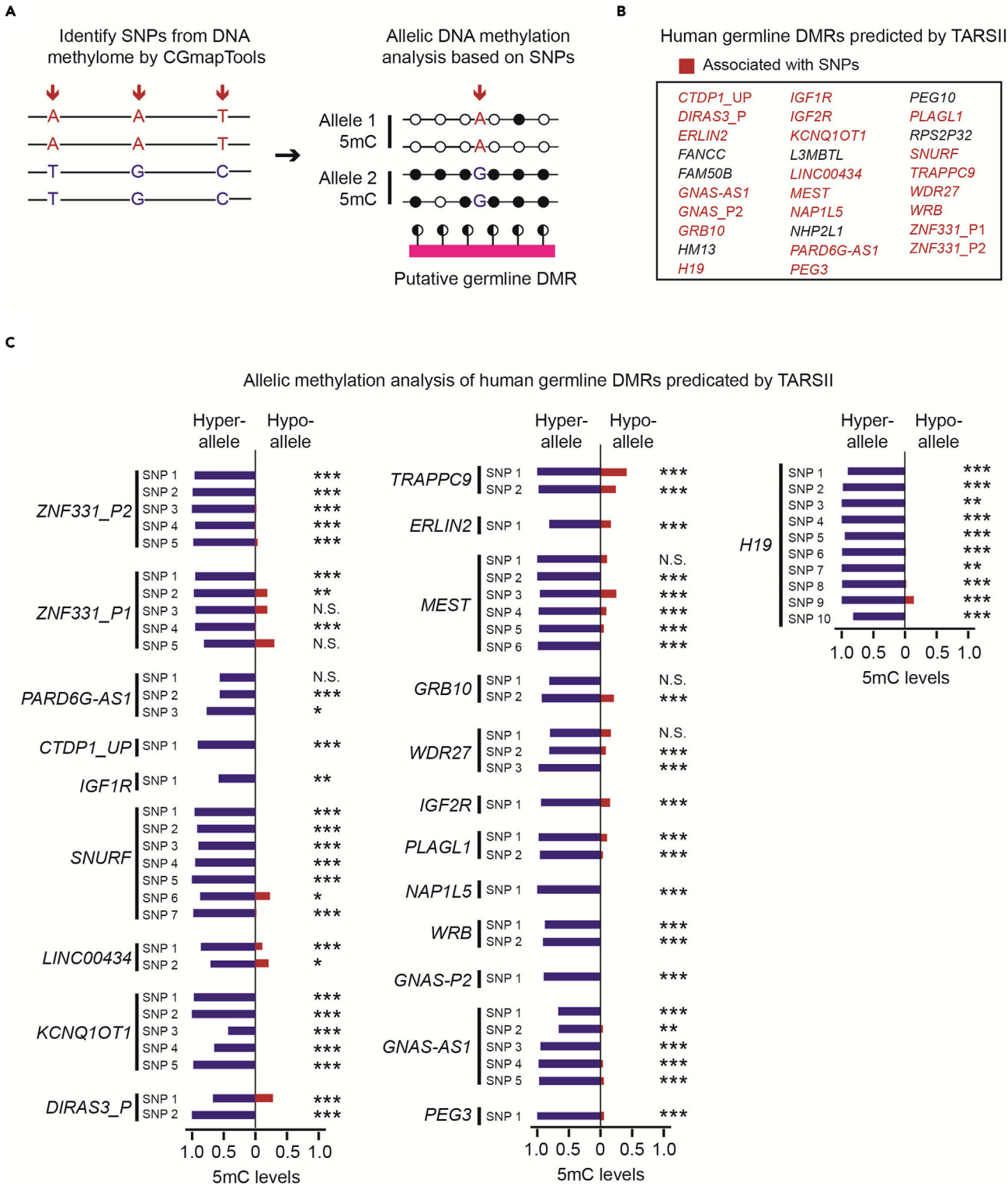


Figure 4. Validation of predicted germline DMRs by CGmapTools

(A) schematic models showing allelic methylation analysis in predicted germline DMRs by CGmapTools. Red arrows indicate SNPs.

(B) A list of human putative germline DMRs predicted by TARSII. Red color indicates the predicated germline DMRs are associated with SNPs identified from somatic tissue methylomes by CGmapTools.

Figure 4. *Continued*
(C) Bar plots showing allelic DNA methylation of TARSII predicted human germline DMRs. The allelic methylation surrounding each SNP located inside the predicted germline DMRs is calculated by CGmapTools. Stars indicate significance of the allelic methylation differences. *: p-value < 1.0E-3, **: p-value < 1.0E-4, ***: p-value < 1.0E-5. N.S. not statistically significant.

Note: The asm file contains information on allelic DNA methylation value, p-value and false discovery rate for each SNP in the analyzed DNA methylome. A TRUE or FALSE judgement is provided as an inference to users about whether the allelic methylation associated with a particular SNP is significantly different or not (troubleshooting problem 5).

⚠ CRITICAL: Since the length and range of imprinted DMRs predicted by TARSII/CARSII can vary when compared to the real imprinted DMRs, some SNPs located within the predicted imprinted DMRs may actually located outside of the real imprinted DMRs. Thus, to check the accurate location of individual SNP using genome browser tools on UCSC genome browser or integrative genomic viewer (IGV) is highly recommended. In general, informative SNPs should locate within a clear partially methylated domain. On the other hand, it would be more solid if several SNPs within certain DMR are all exhibit significant allelic DNA methylation differences.

16. As an example, we predicted the germline DMRs by TARSII using human somatic tissue methylomes as indicated in "before you begin". In total, 29 of germline DMRs were identified, most of which are maternal germline DMRs except H19 (Figure 4B). 22 of the 29 germline DMRs are associated with at least 1 SNP identified by CGmapTools based on the human somatic tissue methylomes (Figure 4B). All of those SNP-associated germline DMRs exhibit clear allele-specific DNA methylation (Figure 4C), confirming the high accuracy of our approaches.

EXPECTED OUTCOMES

TARSII and CARSII are designed for efficient identification of germline DMRs in animal without the pre-knowledge of SNPs. These approaches can greatly reduce the time and resources needed for imprinting analysis in outbred mammals. TARSII predicts the germline DMRs that are imprinted in different somatic tissues from a genome-wide level. CARSII only focuses on CpG islands and identifies germline DMRs that contain or overlap with CpG islands. Since a single DNA methylome is the minimal requirement for CARSII, tissue-specific germline DMRs can be identified by CARSII. Both TARSII and CARSII allow *de novo* discovery of germline DMRs. For validation of novel germline DMRs predicted by TARSII/CARII, we introduced CGmapTools, which identifies SNPs directly from DNA methylome and calculates allelic DNA methylation associated with those SNPs (Figure 4).

LIMITATIONS

TARSII and CARSII are computational tools for predicting germline DMRs. Due to its SNP-independent nature, TARSII and CARSII can be conveniently applied to germline DMR identification in outbred mammals with minimal requirement of DNA methylomes. However, false discovery rate is unavoidable due to factors such as cell heterogeneity, sequencing depth, and experimental variations. The FDR is higher in CARSII considering only a single DNA methylome is used. To reduce the FDR, one strategy is to perform allelic DNA methylation analysis using CGmapTools, to validate the predicted DMRs with SNPs inside (Figures 4A–4C). This validation is especially important if the users attempt to investigate the somatic DMRs predicted by TARSII/CARSII, as the lack of support by inherited allelic DNA methylation from gametes could increase the FDR.

Notably, TARSII and CARSII are not efficient to predict short DMRs (CpG number < 10) because short DMRs are likely to have fewer CpG sites and lower reads coverage. By simply decreasing cutoffs of CpG number and reads coverage will result in significant increase in FDR (Chu et al., 2021). To reduce the FDR in TARSII and CARSII, we included algorithms with intention to remove the regions with low CpG density or low coverage of sequencing reads. Nevertheless, short germline DMRs are

not common and the majority of germline DMRs can be efficiently identified using our approaches in at least mouse, human and monkey (Chu et al., 2021).

TROUBLESHOOTING

Problem 1

TARSII/CARSII does not generate expected file contents after finishing the first step (see steps 2–3 and 9–10).

Potential solution

Please make sure your input files are correctly formatted. Especially, only sam file generated from Bismark and deduplicated with Picard tools (MarkDuplicates) is compatible with TARSII/CARSII. For input file format details, please refer to "materials and equipment" and Figure 1.

Problem 2

The number of germline DMRs predicted by TARSII/CARSII is much fewer than expected (see steps 6–8 and 11–12).

Potential solution

In our protocol we do not specify the requirement of sequencing depth and reads length for TARSII/CARSII prediction, as those can be flexible. However, since TARSII and CARSII are reads-based approaches, deeper in sequencing depth and longer in reads length for the DNA methylomes used will improve the prediction outcome. For a general recommendation, over 100 bp in reads length and over 100 million monoclonal reads for each DNA methylome will ensure a reasonable outcome for TARSII/CARSII.

Problem 3

Some predicted germline DMRs by TARSII from one group of datasets are not identified as germline DMRs in another group of datasets (see steps 6–8).

Potential solution

In outbred mammals, some imprinted regions can be influenced by SNPs or other variations in DNA sequences from different individuals. Thus, it is possible that the predicted germline DMRs are individual-specific. To validate those germline DMRs, analysis of allelic DNA methylation is required. To better reduce the influences of individual genetic background on germline DMR prediction, we highly recommend including somatic tissues from different individuals in TARSII.

Problem 4

The germline DMRs predicted by CARSII do not meet the expected accuracy (see steps 11–12).

Potential solution

Since only a single DNA methylome is applied in CARSII, the predicted results can vary among different DNA methylomes. To increase the performance of CARSII, we recommend users to apply a few DNA methylomes from different sources if available. By comparing results from different DNA methylomes of the same tissue, users can select out a list of common predicted germline DMRs, which will be more accurate.

Problem 5

Significant differences of allelic DNA methylation can be observed on certain SNPs from the calculated results of CGmapTools, but CGmapTools provides a "FALSE" judgement (see step 15).

Potential solution

CGmapTools sets a relatively high standard to exclude the SNPs with potential bi-allelic methylation states, which sometimes will also exclude the SNPs within the known imprinted regions. To help

make an accurate judgement, a few more SNPs in the same predicted germline DMRs should be included in analysis. If available, bisulfite PCR including the targeted SNP can be performed to provide extra information, considering the genomic regions associated with a certain SNP by bisulfite PCR can be much longer than that by the sequencing reads.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yi Zhang (yzhang@genetics.med.harvard.edu)

Materials availability

This study did not generate new unique reagents.

Data and code availability

This study did not generate any unique datasets.

To help the users get started, we listed several public datasets of DNA methylomes in human, monkey and mouse with accession numbers provided in key resources table.

The TARSII and CARSII packages are available for download at https://doi.org/10.5281/zenodo.5484230. Additional script for integration analysis of CGmapTools and TARSII/CARSII in this protocol is available through supplemental information or Mendeley Data (https://doi.org/10.17632/747r4k4mnz.1)

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xpro.2022.101240.

ACKNOWLEDGMENTS

This work is supported by NIH (R01HD092465). Y.Z. is an investigator of the Howard Hughes Medical Institution.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.Z.; methodology, W.Z.; data analysis, W.Z.; writing, W.Z. and Y.Z.; supervision, Y.Z.; funding acquisition, Y.Z.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Babak, T., DeVeale, B., Tsang, E.K., Zhou, Y., Li, X., Smith, K.S., Kukurba, K.R., Zhang, R., Li, J.B., van der Kooy, D., et al. (2015). Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. Nat. Genet. *47*, 544–549.

Barlow, D.P., and Bartolomei, M.S. (2014). Genomic imprinting in mammals. Cold Spring Harb Perspect. Biol. *6*, a018382.

Chen, Z., and Zhang, Y. (2020). Maternal H3K27me3-dependent autosomal and X chromosome imprinting. Nat. Rev. Genet. *21*, 555–571.

Chu, C., Zhang, W., Kang, Y., Si, C., Ji, W., Niu, Y., and Zhang, Y. (2021). Analysis of developmental imprinting dynamics in primates using SNP-free methods to identify imprinting defects in cloned placenta. Dev. Cell *56*, 2826–2840.e7.

Court, F., Tayama, C., Romanelli, V., Martin-Trujillo, A., Iglesias-Platas, I., Okamura, K., Sugahara, N., Simon, C., Moore, H., Harness, J.V., et al. (2014). Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. Genome Res. *24*, 554–569.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

Guo, W., Zhu, P., Pellegrini, M., Zhang, M.Q., Wang, X., and Ni, Z. (2018). CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. Bioinformatics *34*, 381–387.

Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D., and Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat. Genet. *45*, 1198–1206.

Inoue, A., Jiang, L., Lu, F., Suzuki, T., and Zhang, Y. (2017). Maternal H3K27me3 controls DNA methylation-independent imprinting. Nature *547*, 419–424.

Joshi, R.S., Garg, P., Zaitlen, N., Lappalainen, T., Watson, C.T., Azam, N., Ho, D., Li, X.,

Antonarakis, S.E., Brunner, H.G., et al. (2016). DNA methylation profiling of uniparental disomy subjects provides a map of parental epigenetic bias in the human genome. Am. J. Hum. Genet. *99*, 555–566.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics *27*, 1571–1572.

Leng, L., Sun, J., Huang, J., Gong, F., Yang, L., Zhang, S., Yuan, X., Fang, F., Xu, X., Luo, Y., et al. (2019). Single-cell transcriptome analysis of uniparental embryos reveals parent-of-origin effects on human preimplantation development. Cell Stem Cell *25*, 697–712.e6.

Wang, L., Zhang, J., Duan, J., Gao, X., Zhu, W., Lu, X., Yang, L., Zhang, J., Li, G., Ci, W., et al. (2014). Programming and inheritance of parental DNA methylomes in mammals. Cell *157*, 979–991.

Zink, F., Magnusdottir, D.N., Magnusson, O.T., Walker, N.J., Morris, T.J., Sigurdsson, A., Halldorsson, G.H., Gudjonsson, S.A., Melsted, P., Ingimundardottir, H., et al. (2018). Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. Nat. Genet. *50*, 1542–1552.