Computational and Structural Biotechnology Journal 23 (2024) 566–576

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

AIMER: A SNP-independent software for identifying imprinting-like allelic methylated regions from DNA methylome

Yanrui Luo [1], Tong Zhou [1], Deng Liu, Fan Wang, Qian Zhao [*]

Department of Cell Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China

ARTICLE INFO

Keywords:
Differentially methylated region
Allele-specific methylation
SNP-independent
Imprinting-like
AIMER

ABSTRACT

Genomic imprinting is essential for mammalian growth and embryogenesis. High-throughput bisulfite sequencing accompanied with parental haplotype-specific information allows analysis of imprinted genes and imprinting control regions (ICRs) on a large scale. Currently, although several allelic methylated regions (AMRs) detection software were developed, methods for detecting imprinted AMRs is still limited. Here, we developed a SNP-independent statistical approach, AIMER, to detect imprinting-like AMRs. By using the mouse frontal cortex methylome as input, we demonstrated that AIMER performs very well in detecting known germline ICRs compared with other methods. Furthermore, we found the putative parental AMRs AIMER detected could be distinguished from sequence-dependent AMRs. Finally, we found a novel germline imprinting-like AMR using WGBS data from 17 distinct mouse tissue samples. The results indicate that AIMER is a good choice for detecting imprinting-like (parent-of-origin-dependent) AMRs. We hope this method will be helpful for future genomic imprinting studies. The Python source code for our project is now publicly available on both GitHub (https://github.com/ZhaoLab-TMU/AIMER) and Gitee (https://gitee.com/zhaolab_tmu/AIMER).

1. Introduction

Genomic imprinting is a vital phenomenon during mammalian growth and development [1,2], which refers to genes preferentially expressed from either paternal or maternal allele [3–5]. In mammals, such imprinting gene expression is regulated by allele-specific methylation (ASM) in some *cis*-acting regulatory regions in almost all known cases [6–8]. Aberrant DNA methylation of allelic methylated regions (AMRs) is associated with certain diseases. For example, hypermethylation of the *H19* promoter region is a major cause of the clinical features of gigantism and/or asymmetry seen in Beckwith-Wiedemann syndrome or isolated hemihypertrophy [9]. Another example is that hypomethylation of the *NNAT* promoter region and hypermethylation in the *IGF2* region are characteristics of Wilms tumor [10]. In addition, the *MKRN3* mutations were identified to lead to precocious puberty [11]. Therefore, identifying imprinted genes and their regulatory mechanisms is essential for understanding mammalian development and aberrant genomic imprinting diseases [12–14].

High-throughput sequencing technology allows the analysis of imprinted genes on a large scale [15,16]. Whole genome bisulfite sequencing (WGBS) is a practical and informative method to study DNA methylome [17,18]. Whole genome bisulfite sequencing (WGBS) methylome accompanied with parental haplotype-specific information allows analysis of imprinted genes and imprinting control regions (ICRs) on a large scale [18,19]. However, the parental haplotype-specific information is often difficult to obtain, and therefore become the bottleneck for genomic imprinting studies.

For decades, several SNP-independent methods for discovering AMRs have been published. Amrfinder applies two statistical models (one assuming that both alleles are equally methylated and the other supposing that the two alleles have distinct methylation statuses) to assess whether the area is ASM by comparing the likelihood of the two models [20]. MethylMosaic (not publicly accessible) employs a bimodal methylation model to identify AMRs [21]. DAMEfinder calculates an ASM score mainly using two strategies: one (the SNP-based strategy) calculates the heterogeneity of a single CpG site, and the other (the tuple-based strategy) calculates the score based on the read count of paired CpG sites [22]. However, since AMRs may result from not only imprinting (parent-of-origin-dependent) but also other factors such as sequence-dependent (different strain backgrounds) and random

* Corresponding author.
E-mail address: zhaoq@tmu.edu.cn (Q. Zhao).
[1] Equal contribution.

https://doi.org/10.1016/j.csbj.2023.12.038
Received 6 October 2023; Received in revised form 23 December 2023; Accepted 23 December 2023
Available online 3 January 2024
2001-0370/© 2024 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

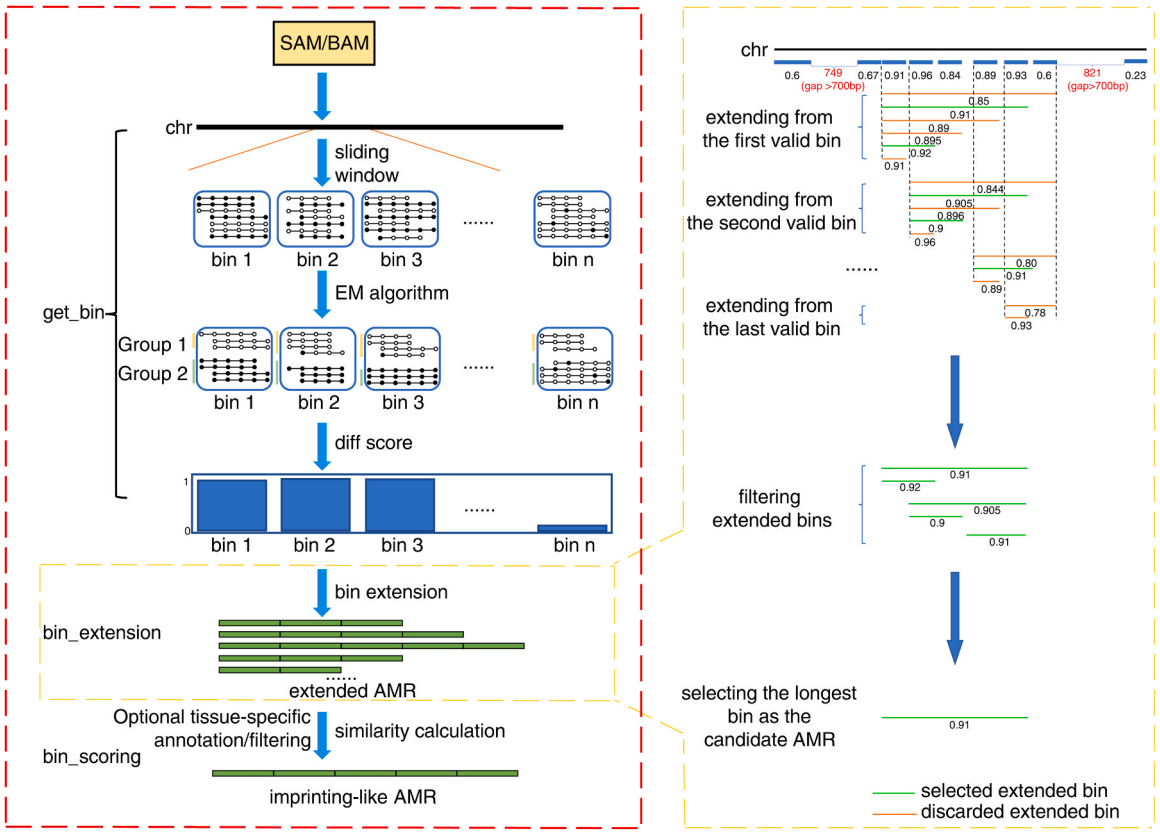


Fig. 1. Overview of AIMER framework. Left. AIMER workflow. The first part is "get_bin", which divides the genome into certain length bins by sliding windows. According to the DNA methylation level of the reads in each bin, reads were divided into two groups with the EM algorithm. Then, the differential methylation scores (diff_score) between the two groups were calculated. The second part is "bin_extension", which means that adjacent bins are joined to obtain a longer region. The third part, "bin_scoring", excludes tissue-specific regions and utilizes logistic regression to estimate the likelihood of being an imprinting-like AMR. **Right.** The bin_extension sketch diagram. Valid bins have a diff_score > 0.85 and a minor composition proportion > 0.3. When the first valid bin is identified, join the adjacent bin within 700 bp. (a valid or invalid bin). Next, from the second valid bin, then repeat the above step until the last valid bin. After that, filter the extended bins with the below conditions: averaged diff_score > 0.9, the max CG count of an included bin > 10, and the averaged proportion of minor composition > 0.4. At last, we select the longest bin as the candidate AMR. The green line represents the selected merged bins that pass the filtering criteria, and the orange line represents the discarded merged bins.

generation [19,21,23–26], none of these approaches are specifically designed to identify imprinted AMRs.

AIMER, our SNP-independent approach for identifying imprinting-like AMRs, is inspired by the concept that DNA methyltransferases always function in a certain region rather than a single CpG site [27–29]. In addition, it used a mixed probability model and logistic regression calculation to ensure the detected regions are imprinting-like. By taking the mouse F1i methylome as input [10], we demonstrated that AIMER performs very well in detecting known germline imprinting control regions (ICRs) compared with other methods. Furthermore, we found AMRs identified by AIMER could be distinguished from sequence-dependent AMRs [19]. Finally, we applied the model to the methylome of 17 different healthy mouse tissues and discovered a novel germline imprinting-like AMR. Our results indicate AIMER is a good choice for parent-of-origin-dependent (imprinting) AMR detection, and our method will provide convenience for future mammalian development and disease studies.

2. Method

2.1. Overview

In this study, we developed a methylome-based SNP-independent method, AIMER, to detect imprinting-like AMRs in a single sample. AIMER, developed in Python 3.9, is currently available for Linux and consists of three sub-commands: "get_bin", "bin_extension", and "bin_scoring". We assumed that, in pure cell populations, the CpG methylation levels of WGBS reads in a certain genomic interval are consistent (either hypermethylated or hypomethylated), except for reads in AMRs. If (i) WGBS reads in a given genomic interval could be classified into two distinct groups (one hypermethylation group and one hypomethylation group), and (ii) the read count in the hypermethylation group and hypomethylation group are similar, we designate the interval as a candidate bin. Considering that in most cases, (i) a practical sample consists of different cell populations (one major component and many other minor components), and (ii) ASM was caused not only by imprinting but also by strain-specific and randomly generated factors. AIMER uses imprinting similarity score calculation and optional tissue-specific DMR filtering to ensure the reliability of imprinting-like AMRs' identification.

2.2. Data source

The WGBS data used for the analysis in this study was obtained from the GEO public database. Data used for simulation tests and performance comparison with other methods are available under GEO accession number GSE33722 [19]. In the study, Bing Ren and colleagues performed reciprocal crosses between two inbred mouse strains, 129 × 1/SvJ (129) and Cast/EiJ (Cast) [19]. WGBS was conducted on frontal cortex DNA from adult F1 progenies of the initial cross 129 (female) × Cast (male) (denoted as F1i) and the reciprocal cross Cast (female) × 129 (male) (denoted as F1r), generating 1.54 billion (F1i) and 1.33 billion

Table 1
Known imprinted DMRs in mouse. Genomic coordinates based on UCSC Genome Browser, Jul. 2007 release, NCBI37/mm9; M, maternal; P, paternal; Known germline DMRs are marked by "Y", "N" denotes non-germline DMRs, and "NA" indicates not available.

Locus	Chr	Start	End	Length (bp)	Meth-allele	Germline imprints?	References
Gpr1/Zdbf2	chr1	63,296,857	63,327,099	30,243	P	Y	Kobayashi et al., 2009; Hiura et al., 2010
Mcts2/H13	chr2	152,512,010	152,512,663	654	M	Y	Wood et al., 2007
Nesp	chr2	174,109,010	174,113,395	4386	P	N	Peters et al., 1999; Kelsey et al., 1999; Mita et al., 2009
Nespas/ Gnasxl	chr2	174,118,404	174,125,287	6884	M	Y	Kelsey et al., 1999; Coombes et al., 2003; Chotalia, Mita et al., 2009
Gnas1a	chr2	174,152,611	174,153,503	893	M	Y	Liu et al., 2000; Williamson et al., 2004; Liu et al., 2005; Mita et al., 2009
Nnat	chr2	157,385,088	157,387,520	2432	M	Y	Schulz et al., 2009; Xu, Yuxin et al., 2016
Peg10/Sgce	chr6	4696,303	4699,370	3068	M	Y	Ono et al., 2003
Mest (Peg1)	chr6	30,685,840	30,689,965	4126	M	Y	Lefebvre et al., 1997; Lucifero et al., 2002; Lucifero et al., 2004
Herc3/ Nap1l5	chr6	58,857,395	58,857,788	394	M	Y	Smith et al., 2003;Wood et al., 2007
Peg3/Usp29	chr7	6680,067	6685,920	5854	M	Y	Li et al., 2000; Lucifero et al., 2002; Kim et al., 2003
Snurf/Snrpn	chr7	67,148,026	67,150,169	2144	M	Y	Shemer et al., 1997; Bielinska et al., 2000
Ndn	chr7	69,493,100	69,493,181	82	M	N	Hanel and Wevrick, 2001
Magel2	chr7	69,521,307	69,522,167	861	M	NA	Dindot et al., 2009; Sharp et al., 2010
Mkrn3	chr7	69,564,012	69,565,740	1729	M	N	Hershko et al., 1999
Peg12	chr7	69,608,471	69,609,019	549	M	NA	Chai et al., 2001
Inpp5f	chr7	135,831,638	135,831,747	110	M	Y	Choi et al., 2005; Wood et al., 2007
H19 promoter	chr7	149,763,483	149,765,230	1748	P	N	Kimberly et al., 1997
H19 ICR	chr7	149,765,791	149,767,931	2141	P	Y	Bartolomei et al., 1993; Ferguson-Smith et al., 1993; Tremblay et al., 1995; Thorvaldsen et al., 1998; Ueda et al., 2000
Kcnq1ot1	chr7	150,480,736	150,482,006	1271	M	Y	Engemann et al., 2000; Fitzpatrick et al., 2002; Yatsuki et al., 2002
Cdkn1c	chr7	150,645,240	150,647,381	2142	P	N	Bhogal et al., 2004
Cdkn1c upstream	chr7	150,649,567	150,649,883	317	P	-	Yatsuki et al., 2002, Genome Res.; Lewis et al., 2004, Nat. Genet.
Rasgrf1	chr9	89,771,945	89,780,072	8128	P	Y	Plass et al., 1996; Shibata et al., 1998; Yoon et al., 2002
Plagl1	chr10	12,809,928	12,812,145	2218	M	Y	Smith et al., 2002
Grb10	chr11	11,923,325	11,926,800	3476	M	Y	Arnaud et al., 2003; Hikichi et al., 2003; Shiura et al., 2009
Zrsr1/ Commd1	chr11	22,871,545	22,874,145	2601	M	Y	Hayashizaki et al., 1994; Shibata et al., 1997; Wood et al., 2007; Schulz et al., 2009; Joh, Keiichiro et al., 2018
Dlk1	chr12	110,697,919	110,700,243	2325	P	N	Takada et al., 2002;
Dlk1-Gtl2 IG	chr12	110,763,965	110,768,609	4645	P	Y	Takada et al., 2002; Lin et al., 2003
Gtl2	chr12	110,777,813	110,781,249	3437	P	N	Takada et al., 2002;
Peg13/ Trappc9	chr15	72,632,245	72,641,614	9370	M	Y	Smith et al., 2003; Ruf et al., 2007
Slc38a4	chr15	96,884,431	96,886,172	1742	M	Y	Smith et al., 2003; Chotalia et al., 2009
Airn/Igf2r	chr17	12,934,626	12,935,815	1190	M	Y	Stoger et al., 1993; Wutz et al., 1997
Igf2r	chr17	12,962,643	12,962,696	54	P	N	Stoger et al., 1993; Wutz et al., 2001
Impact	chr18	13,131,356	13,133,257	1902	M	Y	Okamura et al., 2000

(F1r) uniquely mapped reads. To distinguish the parental origin information, 20.4 million SNPs between the 129 and Cast genomes were identified. In addition, we downloaded previously published 17 mouse WGBS datasets under the GEO accession number GSE42836 to validate the model and investigate AMRs in normal mouse tissues [30]. The 17 mouse tissues include bone marrow, cerebellum, colon, cortex, heart, intestine, kidney, liver, lung, olfactory bulb, pancreas, placenta, skin, spleen, stomach, thymus, and uterus. SRX11551224 was employed to conduct performance tests with different read lengths. The sperm and oocyte data used in the article are accessible under GEO accession number GSE56697 [31].

2.3. The probabilistic model construction

In a given genomic interval (CG count >= 5, in a 300 bp bin by default), m_1 and m_2 are denoted as CpG methylation level of the two classified groups, respectively, and their corresponding proportions are α_1 *and* α_2. The two classified groups consist of two different methylated reads: the hyper- and hypo-methylated reads. Since $\alpha_1 + \alpha_2 = 1$, the methylation pattern in the interval could be modeled as $\Theta = (m_1, m_2, \alpha_1)$. Let X be a set of WGBS reads in the interval, and x be a read from X. l_x is denoted as the count of CpGs, then $x_i = 1$ if the i-th CpG in x is methylated, and $x_i = 0$ otherwise, where $i = 1, 2, \ldots, l_x$.

M_x and U_x are denoted as the number of methylated and unmethylated CpGs in read x, respectively.

$$M_x = \sum_{i=1}^{l_x} x_i \quad U_x = \sum_{i=1}^{l_x} (1 - x_i)$$

The probability of observing read x belongs to the j-th group ($j = 1$, 2) is:

$$p_{x,j} = \prod_{i=1}^{l_x} \left(m_j + (1 - m_j)(1 - x_i)\right) = m_j^{M_x} \bullet (1 - m_j)^{U_x}$$

As the sequence x may come from either group 1 or group 2, the probability of observing x is:

$$p(x) = \alpha_1 p_{x,1} + \alpha_2 p_{x,2}$$

So, the probability of observing the set of reads in the genomic interval X is:

$$p(X) = \prod_{x \in X} p(x)$$

To sum up, given a set of bisulfite reads in a genomic interval, m_1, m_2, and α_1 can be estimated by maximizing the log-likelihood of $p(X)$.

$$\widetilde{\Theta} = \arg\max_{\Theta} l(x) = \arg\max_{\Theta} \prod_{x \in X} p(x) = \arg\max_{\Theta} \sum_{x \in X} \log p(x)$$

Expectation-Maximization (EM) is applied to solve the optimization problem. EM algorithm starts with the randomly selected parameters

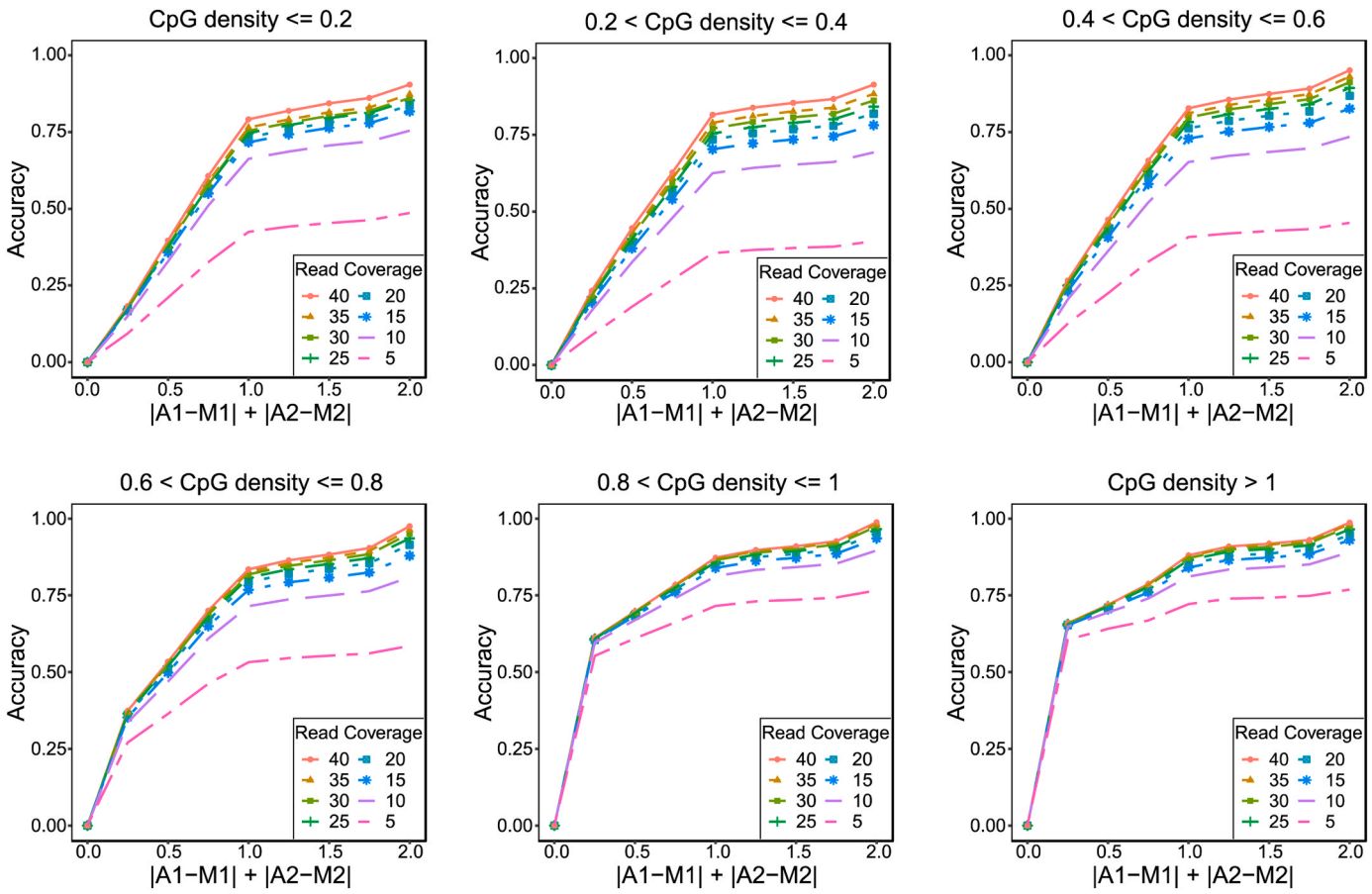


Fig. 2. The accuracy curves of AIMER in different depths and CpG densities. A1 and A2 represent the observed methylation levels of homologous chromosomes in a certain bin; M1 and M2 represent estimated methylation levels for the same bin. Accuracy = Positive bin / Total bin, CpG density = [number of CpGs / (number of Cs) * number of Gs] * length of the region in nucleotides.

$\Theta = (m_1, m_2, \alpha_1)$. E steps and M steps are repeated recursively until converge to a local maximum of log-likelihood function.

The final EM algorithm can be formulated as:

E-step:

$$Q_x(j) = p(z_x = j|x;\Theta) = \frac{\alpha_j p_{x,j}}{\alpha_1 p_{x,1} + \alpha_2 p_{x,2}}$$

M-step:

$$\begin{cases} m_j = \frac{\sum_{x\in X} Q_x(j) M_x}{\sum_{x\in X} Q_x(j) l_x}, j = 1,2 \\ \alpha_1 = \frac{\sum_x Q_x(1)}{|X|} \end{cases}$$

Furthermore, diff_score is calculated by | m1 - m2 | to represent the differential methylation level between two classified groups in each bin (Fig. 1 and Supplementary Fig. S1).

2.4. AMRs identification

As depicted in Supplementary Fig. S1, the diff_score is calculated using the probabilistic model for each bin. In the next step (Supplementary Fig. S1), valid bins (with diff_score > 0.85 and proportion of minor composition > 0.3) are aligned with other adjacent bins if the distance between them is less than 700 bps, the averaged diff_score is greater than 0.9, the maximum CG count of the included bins is greater than 10, and the averaged proportion of minor composition is greater than 0.4 after the extension.

2.5. AMRs scoring

The extended candidate AMRs could be classified into allelic and non-allelic origins. The allelic-origin AMRs occur when DNA methylation patterns are asymmetrical between alleles (with sequence-dependent, parent-of-origin-dependent, or randomly generated), whereas non-allelic-origin AMRs are caused by other mechanisms, such as tissue or cell type-specific methylation [32]. In order to exclude the non-allelic-origin AMRs, AIMER provides optional parameters for annotating or excluding AMRs that may result from tissue- or cell-type-specific methylation. Mouse and human tissue-specific gene annotation files used in this study were obtained from the Mouse MSigDB and Human MSigDB Collections (https://www.gsea-msigdb.org/gsea/msigdb). Users can also choose their own tissue or cell type-specific gene list for this annotation step. Then, we assigned a score to each candidate AMR to determine its likelihood of being an imprinting-like AMR (Supplementary Fig. S1). The likelihood was calculated using a logistic regression model that included the AMR's diff_score, length, number of CGs, and reads fraction of minor composition.

2.6. Simulation datasets

We randomly extracted reads from F1i mouse methylome under different sequencing depths and CpG densities to test the model's accuracy in different conditions. Additionally, since the initial cross F1i was denoted as 129 (female) × Cast (male), and the reciprocal cross F1r was denoted as Cast (female) × 129 (male), we simulated the paternal methylome with F1i Cast + F1r 129 methylome reads and maternal

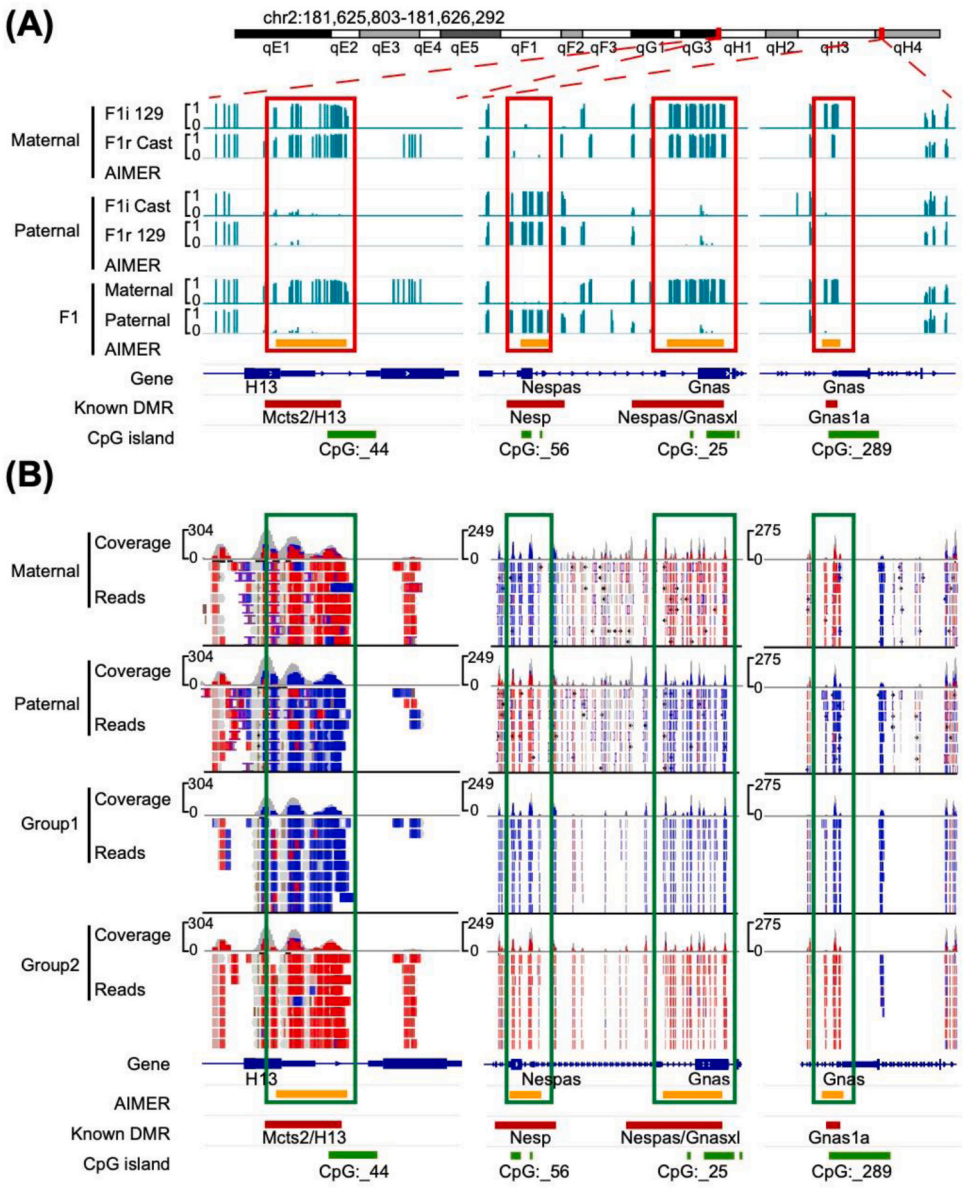


Fig. 3. AIMER performance by using simulated datasets as input. (A) Within "maternal" block, the averaged CGs' methylation level were calculated by F1i 129 and F1r Cast reads, respectively. Within "paternal" block, the averaged CGs' methylation level were calculated by F1i Cast and F1r 129 reads, respectively. Within "F1″ block, the averaged CGs' methylation level were calculated by all of the maternal and paternal reads, respectively. "AIMER" track means the imprinting-like AMRs found by our model when using maternal/paternal/F1 methylome as input. **(B)** Representing reads and reads coverage in each block. "Maternal" block represents reads annotated as F1i 129 and F1r Cast; "paternal" block represents reads annotated as F1i Cast and F1r 129 reads. Reads in Groups 1 and 2 are the classification results by AIMER (input is F1 methylome). The blue bar represents unmethylated CpG, and the red bar represents methylated CpG in a read. Known DMR track refers to the known imprinted DMRs. The "AIMER" track means the imprinting-like AMRs found by our model when using F1 methylome as input.

methylome with F1i 129 + F1r Cast methylome reads. F1 methylome was mixed by F1i + F1r methylome reads. Notably, most of methylome reads naturally contain no heterozygous SNP information in the study [19]. Therefore, F1 methylome contains many unassigned reads (which are useless in the SNP-dependent method) compared to paternal and maternal methylome. The unassigned reads could further facilitate AIMER in effectively detecting AMRs at known imprinted differential methylated regions (or known imprinted DMRs) (Table 1). In addition, separate tests were performed on the SRX11551224 using sequences of 50, 100, and 150 base pairs to evaluate performance under different read lengths.

2.7. P-AS score and S-AS score

The p values derived from the Fisher's exact test were used to calculate an "allele-specific score" (AS score, -log10(*P*-value)) for each CG, which represents the DNA methylation bias for either the parent of origin (P-AS score) or the strain background (S-AS score), and the mean was used to aggregate the CG for each DMR[33]. Positive and negative values were assigned to indicate maternal and paternal preferences for the P-AS score, and 129 and Cast preferences for the S-AS score, respectively. For the P-AS score, the Fisher test was used to calculate an "allele specificity score" (AS score, -log10 (*P*-value)) for each CG to reflect the DNA methylation bias for the parent of origin (P-AS score), and the mean was used to aggregate the CG for each DMR [19]. For the S-AS score, each CG was assessed by Fisher's exact test using

Table 2
Imprinting-like AMR found by AIMER in F1i methylome data. Diff score, differential methylation level between the two groups; Length, the length of the detected imprinting-like AMR; GC, the GC count of the detected AMR; Ratio, the reads proportion of the minimum group; Prob, the probability of AMR; Gene, AMR located in TSS+ /- 2k; Type, the type of known DMR; gDMR, germline DMR; sDMR, somatic DMR; NA, not available.

Chr	Start	End	Diff score	Length	CG	Ratio	Prob	Gene	Known	Type
chr11	11,925,236	11,927,043	0.95	1807	45	0.45	1	*Grb10*	*Grb10*	gDMR
chr11	22,871,795	22,874,209	0.97	2414	25	0.46	1	*Zrsr1,Commd1*	*Zrsr1/Commd1*	gDMR
chr12	110,778,034	110,781,358	0.92	3324	14	0.45	1	*Mir1906-1, Mir1906-2, Meg3*	*Gtl2*	sDMR
chr15	72,638,743	72,641,754	0.96	3011	22	0.43	1	*Peg13*	*Peg13/Trappc9*	gDMR
chr18	13,130,553	13,133,267	0.97	2714	28	0.45	1	*Impact*	*Impact*	gDMR
chr2	157,385,088	157,387,520	0.92	2432	32	0.46	1	*Nnat*	*Nnat*	gDMR
chr2	174,108,943	174,114,716	0.91	5773	31	0.42	1	*Gnas*	*Nesp*	sDMR
chr2	174,119,919	174,126,599	0.93	6680	27	0.45	1	*Gnas,Gnasas1*	*Nespas/Gnasxl*	gDMR
chr2	174,152,514	174,154,940	0.92	2426	33	0.4	1	*Gnas*	*Gnas1a*	gDMR
chr6	30,685,113	30,689,339	0.92	4226	36	0.45	1	*Mest,Mir335*	*Mest(Peg1)*	gDMR
chr6	4696,112	4699,434	0.96	3322	26	0.46	1	*Peg10,Sgce*	*Peg10/Sgce*	gDMR
chr7	6679,962	6684,779	0.95	4817	19	0.45	1	*Usp29,Peg3*	*Peg3/Usp29*	gDMR
chr7	135,831,129	135,833,266	0.95	2137	25	0.43	0.99	*Inpp5f*	*Inpp5f*	gDMR
chr7	150,480,838	150,482,646	0.97	1808	31	0.46	0.99	*Kcnq1ot1*	*Kcnq1ot1*	gDMR
chr7	67,148,067	67,150,207	0.95	2140	16	0.44	0.99	*Snrpn,Snurf*	*Snurf/Snrpn*	gDMR
chr5	111,849,943	111,852,355	0.86	2412	16	0.44	0.98	NA	NA	NA
chr7	149,765,643	149,768,054	0.89	2411	15	0.42	0.98	*Mir675,H19*	*H19* ICR	gDMR
chr10	12,810,085	12,811,893	0.93	1808	30	0.41	0.97	*Plagl1,Hymai*	*Plagl1*	gDMR
chr17	12,934,144	12,935,649	0.97	1505	25	0.46	0.96	*Airn*	*Airn/Igf2r*	gDMR
chr15	96,884,535	96,886,343	0.94	1808	25	0.4	0.95	*Slc38a4*	*Slc38a4*	gDMR
chr6	58,855,907	58,857,429	0.92	1522	29	0.42	0.88	*Nap1l5*	*Herc3/Nap1l5*	gDMR
chr2	152,512,256	152,513,462	0.93	1206	26	0.44	0.7	*Mcts2*	*Mcts2/H13*	gDMR
chr9	89,774,443	89,775,370	0.94	927	29	0.46	0.59	NA	*Rasgrf1*	gDMR
chr13	25,257,278	25,257,892	0.9	614	24	0.48	0.12	NA	NA	NA
chr16	89,897,575	89,898,477	0.89	902	13	0.43	0.07	NA	NA	NA
chr7	142,036,637	142,037,242	0.92	605	22	0.4	0.06	NA	NA	NA
chr12	110,766,230	110,766,831	0.88	601	25	0.44	0.05	NA	*Dlk1-Gtl2* IG	gDMR
chr18	47,880,590	47,880,890	0.96	300	15	0.48	0.05	NA	NA	NA
chr4	136,705,084	136,705,384	0.95	300	18	0.48	0.05	NA	NA	NA
chr17	35,137,177	35,137,477	0.95	300	16	0.45	0.03	*Vars, D17H6S56E-5*	*NA*	NA
chr18	54,858,450	54,859,054	0.93	604	17	0.38	0.03	NA	NA	NA
chr1	40,081,563	40,082,164	0.89	601	14	0.38	0.01	NA	NA	NA

strain-specific reads (129 and Cast) pooled from both F1i and F1r to reflect allelic DNA methylation bias for the strain background (sequence-dependent). -log10 (*P*-value) was then treated on each *P*-value. Sequence-dependent AMRs were aligned by CGs with S-AS score $>= 5$ and merged single CG to S-AS score region (at least 3 or more candidate high S-AS score CGs in a neighboring +/- 2.5 kb window).

We then extracted top 300 sequence-dependent AMRs with the highest S-AS score and compared them to the AMRs identified by AIMER (using only F1i methylome as input, regardless of the strain information). The SNPs information is used to distinguish the allele's parental-of-origin in the progeny strains.

2.8. Methylation processing

The raw methylation sequencing reads were processed using Trimmomatic to remove adaptors and eliminate low-quality reads [34]. The clean reads were then aligned to the mouse reference genome (mm9) and deduplicated using BisMark [35]. The methylation status of each CpG site was extracted from a sorted bam file using the bismark_methylation_extractor function from BisMark.

2.9. Statistical analysis

Receiver operating characteristic (ROC) curve analyses and area under the ROC curve (AUC) computation were performed with the pROC package [36]. True Positive Rate (TPR) and False Positive Rate (FPR) for determining the quality of different methods were also calculated to compare the performance of different methods, as provided below,

$$TPR = TP/(TP + FN)$$

$$FPR = FP/(FP + TN)$$

where T and F denote true/correct and false/incorrect classifications, P and N are the numbers of known positive and negative cases, respectively.

3. Results

3.1. An overview of the AIMER framework

AIMER is a program based on Python and consists of three parts (Fig. 1). In the "get_bin" part, the methylome reads are distributed to different sliding windows. Then, the reads in each sliding window are divided into two groups using the EM algorithm based on their DNA methylation levels. Finally, the difference in methylation between the two groups is calculated (diff_score, Method) for each sliding window. The second part, "bin_extension", connects the individual differentially methylated regions produced in the first step into longer regions under certain criteria (the right panel of Fig. 1). The third step, called "bin_scoring", aims to determine the similarity between the extended region and imprinting AMR, we employ tissue-specific DMRs filtration and logistic regression to identify imprinting-like AMR based on the region's length and CG number, as well as the different levels of DNA methylation and the ratio of sequences from two different sources (Method).

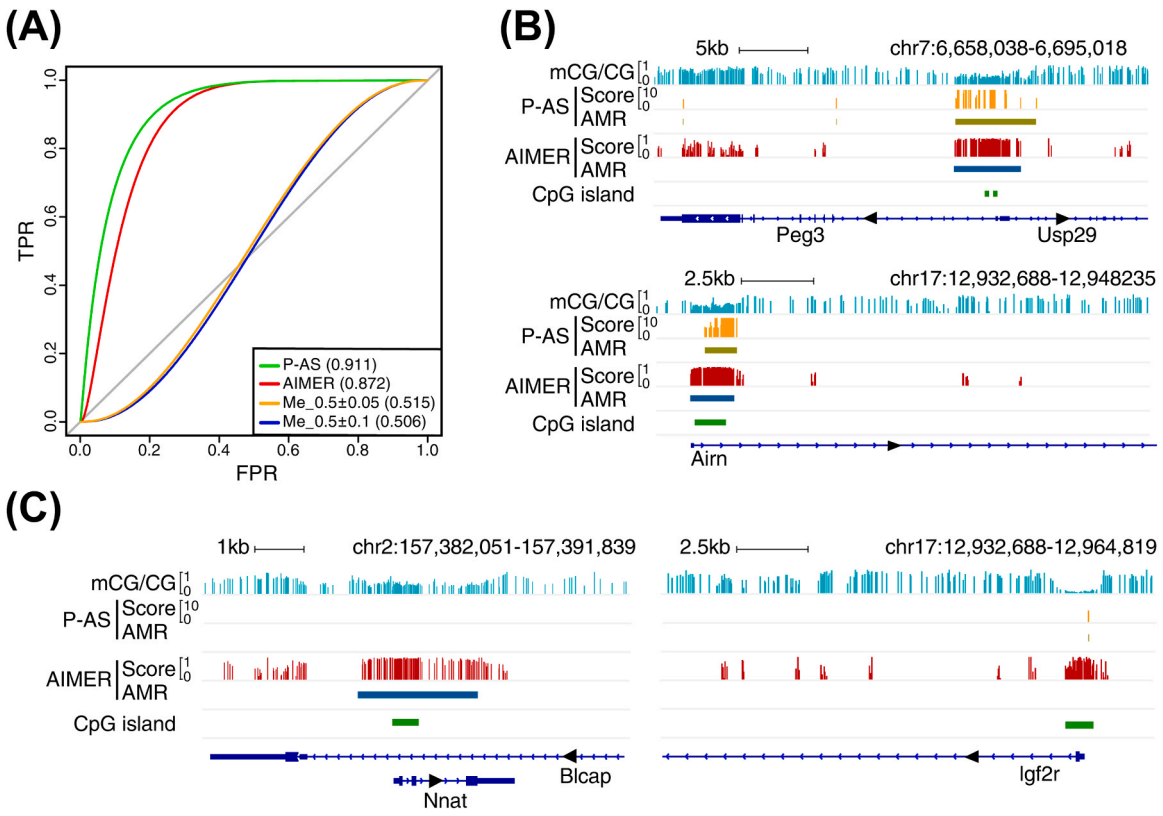


Fig. 4. Evaluation among AIMER and other methods. (A) Performance evaluation by ROC curve, red represents AIMER; green represents P-AS method; orange represents extended bins with averaged methylation level in the range (0.45, 0.55); blue represents extended bins with averaged methylation level in the range (0.4, 0.6). **(B)** Comparison between AIMER and P-AS in known imprinted DMR *Peg3/Usp29* and *Airn*. **(C)** Comparison between AIMER and P-AS in known imprinted DMR *Nnat/Blcap* and somatic imprinting gene areas *Igf2r*. The AMR found by P-AS approach is called P-AS AMR, and the AMR determined by our model is called AIMER AMR. The P-AS score is Fisher's exact test score, and the AIMER score is the differential levels of two groups calculated by our model.

3.2. Simulation

The performance of our model was tested with different sequencing depths and CpG densities. The following results from the Accuracy Curve indicate: 1) the model has a poor predictive power for regions with low CpG density, whereas it performs very well within high CpG density regions (*e.g.*, ICRs); 2) In general, a satisfactory prediction performance may be achieved at a coverage depth of roughly 10X, and sequencing depths more than 15X barely increase prediction accuracy (Fig. 2). Meanwhile, we conducted separate tests on SRX11551224 using sequences of 50, 100, and 150 base pairs to assess software performance, taking the known DMRs as true. The model exhibits outstanding performance for both short and long sequence reads. Additionally, we have noticed a slight performance enhancement with increased read length. It suggests that the model performance improves as the read length increases. (Supplementary Fig. S2).

To investigate whether the model can detect the known imprinted DMRs in mouse, we simulated with chromosome 2 from 3 mouse datasets, including only paternal methylome (F1i Cast and F1r 129), only maternal methylome (F1i 129 and F1r Cast), and mixed F1 (F1i + F1r) data, respectively (Method). When we used only the paternal or maternal methylome as input, the model was unable to detect any known imprinted DMRs. However, as expected, our model was able to successfully identify the known imprinted DMRs when mixed methylome (F1) was used (Fig. 3A). Further analysis (Fig. 3B) indicates that our method can correctly classify the F1 methylome reads into two groups (in other words, reads in group 1 and group 2 are identical to the reads that be annotated as maternal and paternal) without parental information within known imprinted DMRs.

To sum up, our simulation study indicates: AIMER performed well in regions with high CpG density and at roughly 10X depth; it was also able to successfully identify AMRs and correctly classify F1 methylome reads within known imprinted DMRs.

3.3. Evaluation

To evaluate the performance of AIMER, we used F1i methylome (Method) as input to compare AIMER and other methods with the area under the curve (AUC) of the ROC by the pROC package [36]. 32 AMRs were identified by AIMER (Table 2, Method). The evaluation was conducted among four approaches: i) AIMER (our method, using F1i methylome as input), ii) P-AS score (SNP-based cross-hybridization approach, the inputs included: F1i, F1r methylome, parental-SNP information, Method), iii) – iv) by the simple and intuitive way, which directly defined the imprinting-like AMR as the bins with averaged methylation levels in the range of 0.45–0.55 or 0.4–0.6, after bin extension.

The intersection of AMRs discovered by the P-AS score and known imprinted DMRs in mouse was considered true (Table 1 and Supplementary Table S3). The ROC curve indicates that AIMER performs very well in finding parent-of-origin AMRs, which is very close but consistently lower than (or equal to) the P-AS score (Fig. 4A). Further analysis shows that: (1) AIMER could find almost all of the known germline imprinted DMRs; (2) the known germline imprinted DMRs that AIMER identified ranked top in the AIMER result list (Table 2), and their lengths and positions were identical to those found by the P-AS score (Fig. 4B). We should note that *Nnat* was missed by P-AS score selection because of insufficient SNP information (Fig. 4C, left panel). Moreover, by looking into promoters of some P-AS score selected somatic imprinting genes (*e.g.*, *Igf2r*, Fig. 4C, right panel), we found a low averaged methylation level

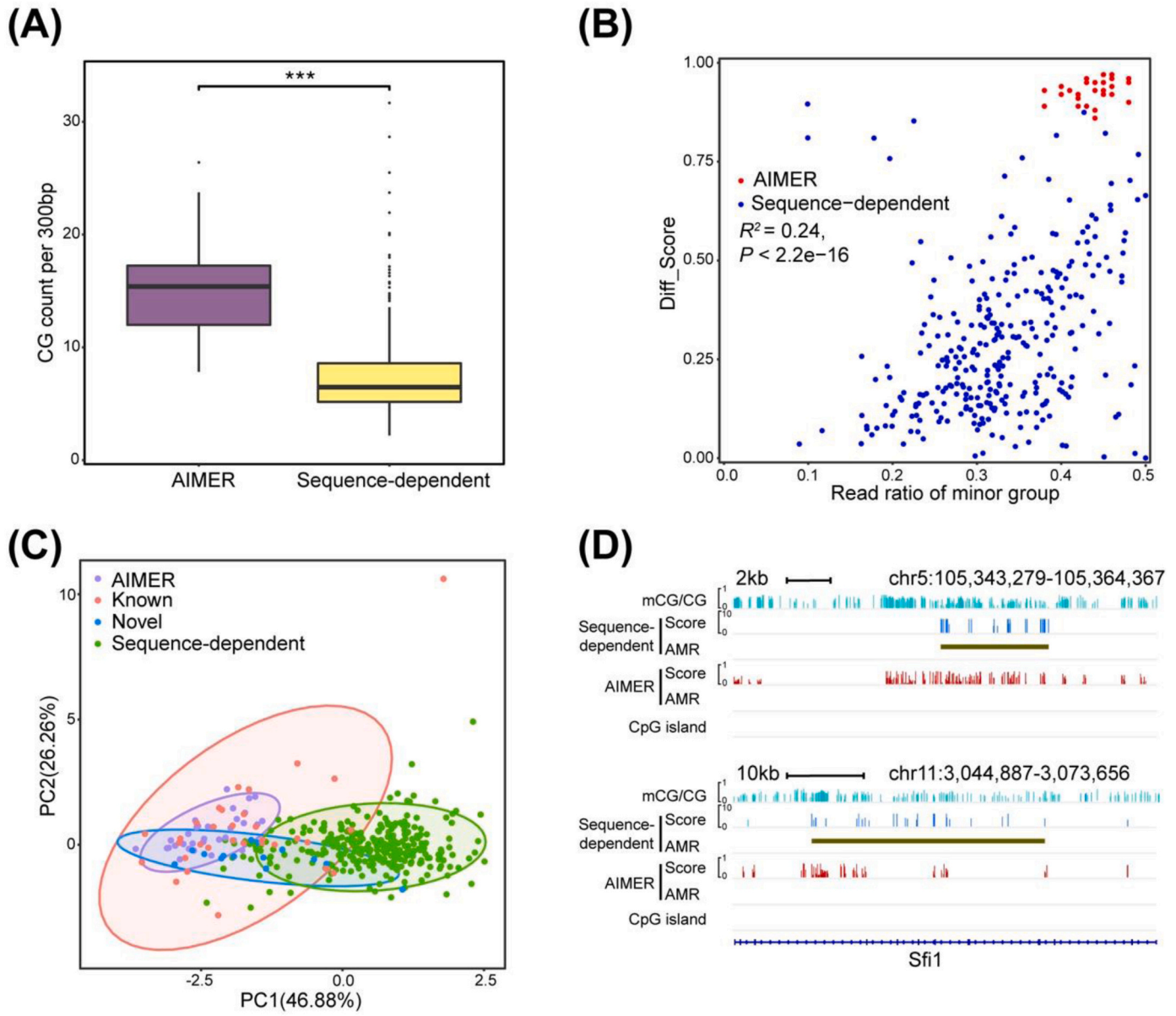


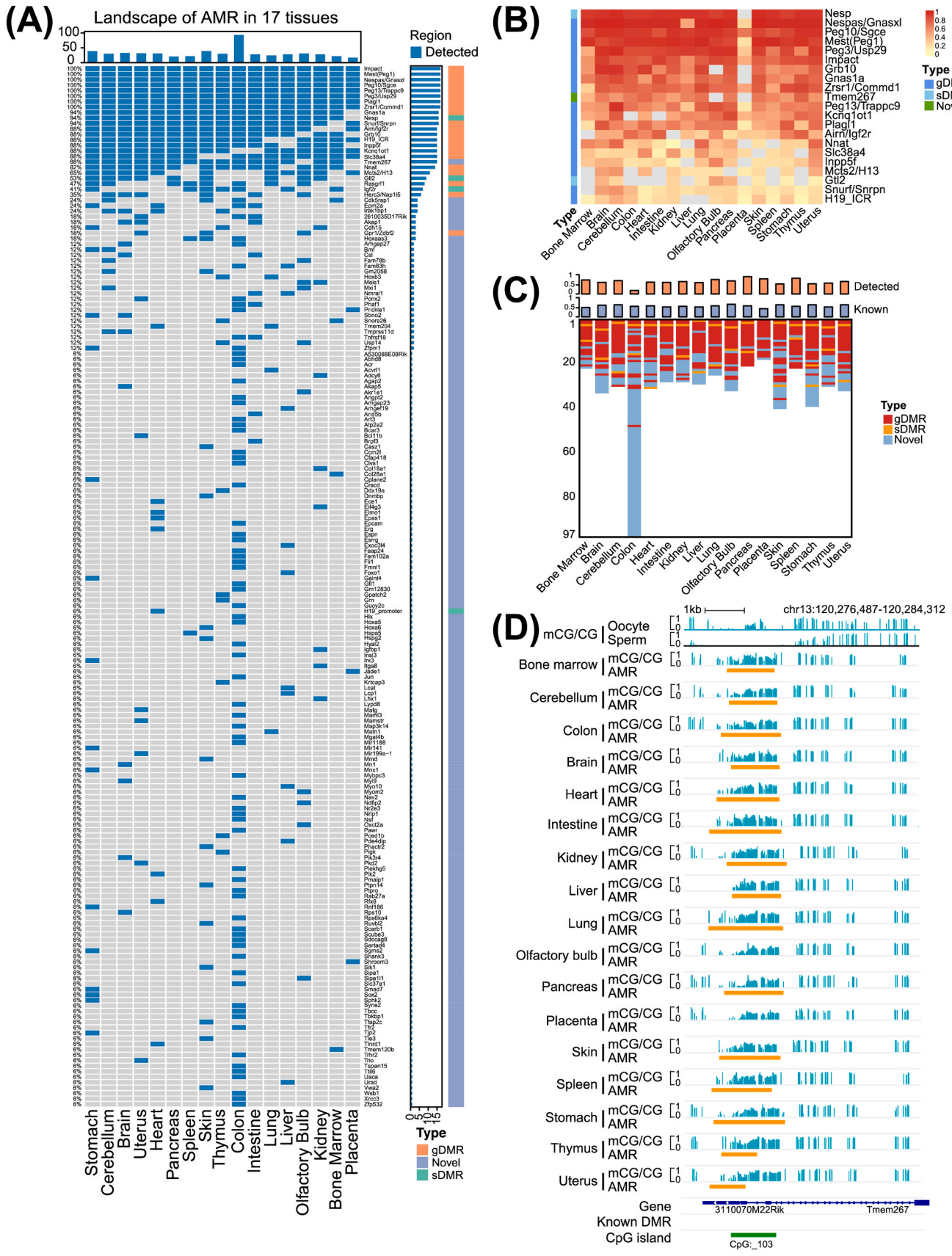
Fig. 5. Comparison of AIMER-discovered AMRs and sequence-dependent AMRs. (A) The box plot compares the CG density of AMRs for AIMER and sequence-dependent. * , $P < 0.05$; * *, $P < 0.01$; * ** , $P < 0.001$; ns, not significant (Wilcoxon test). **(B)** The scatter plot shows the diff_score and read ratio in both AIMER and Sequence-dependent (S-AS) founded AMRs. The ratio is the proportion of reads from the minor group among all reads in a certain bin; diff_score is the differential methylation level of two groups in the same bin. The red dot is the region found by AIMER; the blue dot is the region caused by Sequence-dependent. R^2 indicates the coefficient of determination between the two methods (Pearson correlation). **(C)** PCA plot demonstrated the relationship of AMRs found by AIMER, sequence dependence, and AMRs for both known and novel AMRs found by P-AS score. **(D)** Comparison was made between the AMRs found by AIMER and sequence-dependence. Sequence-dependent AMR is the AMR caused by strain background, while AIMER AMR is the AMR found by our model; Sequence-dependent score results from Fisher's exact test, while the AIMER score is the differential levels calculated by our model between two groups.

(mCG/CG track, calculated by all methylome reads, regardless of carrying SNP information or not), which apparently result in a negative AIMER result. A possible explanation for the P-AS score positive, but AIMER negative results could be the different input datasets for the two approaches: for P-AS score calculation, the input was about 35% of total methylome reads, which means 65% methylome reads were useless because of the deficiency of SNP information [19]; whereas the SNP independent approach AIMER could make full use of methylome reads.

Additionally, we conducted separate comparisons between the outcomes of P-AS and AIMER utilizing F1i and F1r data, respectively (Table 2, Supplementary Table S2 and S3). The results of our study demonstrate that AIMER achieves superior performance, which is extremely close to the P-AS score, regardless of whether we utilize initial or reciprocal data (Supplementary Fig. S3A). Moreover, the majority of the AMRs present in P-AS can be found by AIMER in F1i and F1r (Supplementary Fig. S3B). Hence, our approach may effectively detect imprinting-like AMRs using a single set of methylation data.

Although there is no software specifically designed to detect imprinting-like AMRs, we compared the existing SNP-independent AMR detection software, amrfinder in MethPipe and methtuple in DAMEfinder [20,22]. Supplementary Fig. S4 A shows that AIMER and AMR-detecting tools share a similar AUC. However, the total number of regions detected by AMR-detecting tools is thousands of times greater than the known imprinted DMRs they have found, resulting in very low precision (Supplementary Fig. S4 B and Supplementary Table S1).

In conclusion, AIMER performs as well as the P-AS method (SNP-



(caption on next page)

Fig. 6. Application of AIMER to different tissues. (A) The landscape of AMR in 17 tissues. Each column is a tissue and each row is an AMR. The blue grid represents that tissue can detect a certain AMR; the grey grid represents not detected. The orange grid represents germline DMR; the dusty blue grid represents novel DMR; the green grid represents somatic DMR. "Detected" represents the percentage of known imprinted DMRs among the AIMER-discovered AMRs. **(B)** The heatmap of AMRs found in different tissues. The red to yellow color reflects the similarity between AIMER-discovered AMRs and imprinted DMRs. Grey represents did not find the region in that sample. Only regions that were discovered in at least half of the samples by AIMER are shown on the heatmap. **(C)** Heatmap shows the similarity of AIMER-discovered AMR and imprinted DMR for multiple tissues. The vertical coordinate is the numerical order obtained by sorting the similarity score, the similarity decreases from the top to bottom. Each column represents a sample and each grid represents an AMR. The red grid is known germline DMR, the orange grid is known secondary/somatic DMR, and the blue grid is Novel AMR. Detected represents the percentage of known DMRs among the AIMER-discovered AMRs. Known represents the proportion of discovered known DMRs to all known DMRs. **(D)** AIMER found the novel AMR (*Tmem267*) which can be detected in 15 out of 17 tissues. mCG/CG represents a certain CpG methylation level; AMR is the imprinting-like region found by AIMER.

based cross-hybridization approach). Furthermore, AIMER could discover imprinting AMRs that the P-AS method missed because of the regional SNP information deficiency (*e.g.*, *Nnat*). Finally, by making full use of input methylome reads, the result list generated by AIMER, especially the top list, seems closer to known germline ICRs.

3.4. Comparison with sequence-dependent AMRs

ASM can not only arise in a parent-of-origin-dependent manner but also in a way that is dependent on the sequence context (refers to the AMRs caused by different strain backgrounds) [23,37–39]. Therefore, our next question is whether the novel AMRs that AIMER identified (Table 2) are caused by differences in the underlying sequence. The comparison between AIMER AMRs and sequence-dependent AMRs (calculated by S-AS score, Method, Supplementary Table S4) was conducted. The results show that (1) AMRs identified by AIMER contain significantly more CpGs than sequence-dependent AMRs (Fig. 5A); (2) By calculating the averaged diff_score and averaged ratio (features in AIMER result) of the sequence-dependent AMRs, we found all these regions preserve either poor diff_score or poor averaged ratio (Fig. 5B); (3) Principal Component Analysis (PCA) shows no overlap between AIMER AMRs and sequence-dependent AMRs (Fig. 5C). (4) Top 2 sequence-dependent AMRs (ranked by S-AS score) have extremely low diff_score (Fig. 5D, Supplementary Table S4), indicating a huge difference between AIMER and S-AS score results.

In general, these results illustrate clear feature distinctions between sequence-dependent AMRs and putative parental AMRs. Additionally, the findings suggest that novel AMRs generated by AIMER are more likely to arise in a parent-of-origin-dependent manner, rather than being sequence-dependent. In other words, our method may be a good choice for identifying parent-of-origin-dependent AMRs.

3.5. Identification of imprinting like AMRs in 17 mouse tissues

Our model has shown excellent performance by using mouse frontal cortex methylome as input. We then extended the analysis to 17 different mouse tissues. Fig. 6A shows that most known imprinted DMRs share a high detection frequency among the 17 tissues. Furthermore, almost all of the top-ranked AIMER AMRs are known germline imprinted DMRs (Fig. 6B). It seems that the similarity score, along with the count (*e.g.*, top 30 in the AIMER result list), may provide more reliable imprinting-like AMRs (Fig. 6C, Discussion). Finally, we identified one novel AMR (*Tmem267*) that can be detected in 15 out of 17 mouse tissues, except for the placenta and the olfactory bulb (Fig. 6B). This AMR may be a potential germline ICR (Fig. 6D).

4. Discussion

We developed a SNP-independent computational method called AIMER to detect imprinting-like AMRs. Currently, AIMER is the only bioinformatic software that detects imprinting-like AMRs by simply using methylome as input. Moreover, AIMER shares a similar performance with the SNP-based cross-hybridization approach, which is considered to be the benchmark for detecting imprinting (parent-of-origin) AMRs. And, it would be noted that AIMER is independent of SNPs, therefore it is capable of detecting imprinting-like AMRs that naturally possess no heterozygous SNP information on the sequence. The advantage provides the possibility of discovering more potential imprinting AMRs.

The limitations of AIMER include:

First, AIMER cannot determine whether the detected AMRs are maternally or paternally methylated. In the future, we plan to provide an optional function that users could upload parental-specific heterozygous SNP information to help with the determination. Therefore, to help users further understand the result list, we annotated the output regions as tissue-specific if their associated genes are determined as tissue-specific by MSigDB.

Second, AIMER used a mixed probability model and logistic regression calculation to ensure the detected regions are imprinting-like. Although AIMER performs well in imprinting-like AMRs detection, the result list may still be mixed with some unexpected tissue or cell-type specific DMRs. Theoretically, AIMER could identify not only imprinting-like AMRs but also DMRs from (i) the sample analyzed consists of two components (or cell populations) and the proportions of the two components happen to approach 1:1; (ii) the sample analyzed consists of several components (or cell populations), in a certain genomic interval the combination of components could be grouped into two distinct methylation patterns, and the proportion of the two patterns happens to approach 1:1. Therefore, to help users further understand the result list, we annotated the output regions as tissue-specific if their associated genes were determined as tissue-specific by MSigDB.

Additionally, as mentioned in the result part, the similarity score accompanying the rank (*e.g.*, top 30 in the AIMER result list) may provide more reliable imprinting-like AMRs. A multiple-sample experimental design (biological replicate) would also eliminate the AMRs generated by random or other unknown reasons.

AIMER treats any single-base resolution methylome as input (*e.g.*, WGBS, reduced representation bisulfite sequencing, or RRBS). However, it is obvious that the searching area of AIMER would be significantly narrowed down if RRBS data is used as input.

The known imprinted DMRs in human have been collected and summarized in Supplementary Table S5.

CRediT authorship contribution statement

Yanrui Luo: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Tong Zhou:** Investigation, Visualization, Data analysis, Writing – original draft, Writing – review & editing. **Deng Liu:** Investigation, Visualization, Data curation. **Fan Wang:** Investigation, Visualization. **Qian Zhao:** Conceptualization, Project administration, Writing – review & editing. All authors have read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Bing Ren and colleagues for sharing their data. Their pioneered work provides a valuable resource for understanding the mechanisms of imprinting and allele-specific gene expression. This study was supported by the National Natural Science Foundation of China (31671365).

Code availability

AIMER is available open source at https://github.com/ZhaoLab-TMU/AIMER and https://gitee.com/zhaolab_tmu/AIMER.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.12.038.

References

[1] Zeng Y, Chen T. DNA methylation reprogramming during mammalian development. Genes 2019;10.
[2] Bartolomei MS, Ferguson-Smith AC. Mammalian genomic imprinting. Cold Spring Harb Perspect Biol 2011;3.
[3] Barlow DP, Bartolomei MS. Genomic imprinting in mammals. Cold Spring Harb Perspect Biol 2014;6.
[4] Schilling E, El Chartouni C, Rehli M. Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. Genome Res 2009;19:2028–35.
[5] Lo CL, Lumeng L, Bell RL, Liang T, Lossie AC, et al. CIS-acting allele-specific expression differences induced by alcohol and impacted by sex as well as parental genotype of origin. Alcohol Clin Exp Res 2018;42:1444–53.
[6] Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet 2011;12:565–75.
[7] Jeziorska DM, Murray RJS, De Gobbi M, Gaentzsch R, Garrick D, et al. DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. Proc Natl Acad Sci USA 2017;114:E7526–35.
[8] Monk D. Deciphering the cancer imprintome. Brief Funct Genom 2010;9:329–39.
[9] Bliek J, Terhal P, van den Bogaard MJ, Maas S, Hamel B, et al. Hypomethylation of the H19 gene causes not only Silver-Russell syndrome (SRS) but also isolated asymmetry or an SRS-like phenotype. Am J Hum Genet 2006;78:604–14.
[10] Hubertus J, Zitzmann F, Trippel F, Muller-Hocker J, Stehr M, et al. Selective methylation of CpGs at regulatory binding sites controls NNAT expression in Wilms tumors. PLoS One 2013;8:e67605.
[11] Seraphim CE, Canton APM, Montenegro L, Piovesan MR, Macedo DB, et al. Genotype-phenotype correlations in central precocious puberty caused by MKRN3 mutations. J Clin Endocrinol Metab 2021;106:1041–50.
[12] Lim DH, Maher ER. Genomic imprinting syndromes and cancer. Adv Genet 2010;70:145–75.
[13] Elhamamsy AR. Role of DNA methylation in imprinting disorders: an updated review. J Assist Reprod Genet 2017;34:549–62.
[14] Jelinic P, Shaw P. Loss of imprinting and cancer. J Pathol 2007;211:261–8.
[15] Soellner L, Monk D, Rezwan FI, Begemann M, Mackay D, et al. Congenital imprinting disorders: application of multilocus and high throughput methods to decipher new pathomechanisms and improve their management. Mol Cell Probes 2015;29:282–90.
[16] Tucci V, Isles AR, Kelsey G, Ferguson-Smith AC, Erice G. Imprinting, genomic imprinting and physiological processes in mammals. Cell 2019;176:952–65.
[17] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009;462:315–22.
[18] Beck S, Rakyan VK. The methylome: approaches for global DNA methylation profiling. Trends Genet 2008;24:231–7.
[19] Xie W, Barr CL, Kim A, Yue F, Lee AY, et al. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell 2012;148:816–31.
[20] Fang F, Hodges E, Molaro A, Dean M, Hannon GJ, et al. Genomic landscape of human allele-specific DNA methylation. Proc Natl Acad Sci USA 2012;109:7332–7.
[21] Martos SN, Li T, Ramos RB, Lou D, Dai H, et al. Two approaches reveal a new paradigm of 'switchable or genetics-influenced allele-specific DNA methylation' with potential in human disease. Cell Discov 2017;3:17038.
[22] Orjuela S, Machlab D, Menigatti M, Marra G, Robinson MD. DAMEfinder: a method to detect differential allele-specific methylation. Epigenetics Chromatin 2020;13:25.
[23] Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat Genet 2008;40:904–8.
[24] Hamada H, Okae H, Toh H, Chiba H, Hiura H, et al. Allele-specific methylome and transcriptome analysis reveals widespread imprinting in the human placenta. Am J Hum Genet 2016;99:1045–58.
[25] Lo HS, Wang Z, Hu Y, Yang HH, Gere S, et al. Allelic variation in gene expression is common in the human genome. Genome Res 2003;13:1855–62.
[26] Deng Q, Ramskold D, Reinius B, Sandberg R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 2014;343:193–6.
[27] Jabbari K, Bernardi G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. Gene 2004;333:143–9.
[28] Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009;41:178–86.
[29] Zink F, Magnusdottir DN, Magnusson OT, Walker NJ, Morris TJ, et al. Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. Nat Genet 2018;50:1542–52.
[30] Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet 2013;45:1198–206.
[31] Wang L, Zhang J, Duan J, Gao X, Zhu W, et al. Programming and inheritance of parental DNA methylomes in mammals. Cell 2014;157:979–91.
[32] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
[34] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.
[35] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 2011;27:1571–2.
[36] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform 2011;12:77.
[37] Schalkwyk LC, Meaburn EL, Smith R, Dempster EL, Jeffries AR, et al. Allelic skewing of DNA methylation is widespread across the genome. Am J Hum Genet 2010;86:196–212.
[38] Onuchic V, Lurie E, Carrero I, Pawliczek P, Patel RY, et al. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science 2018;361.
[39] Wang H, Lou D, Wang Z. Crosstalk of genetic variants, allele-specific dna methylation, and environmental factors for complex disease risk. Front Genet 2018;9:695.