

M6Allele: a toolkit for detection of allele-specific RNA N⁶-methyladenosine modifications

Yin Zhang^{1,†}, Lin Tang^{1,†}, Shengyao Zhi^{2,†}, Bosu Hu¹, Zhixiang Zuo³, Jian Ren¹, Yubin Xie^{4,*}, and Xiaotong Luo^{1,5,*}

¹Innovation Center of the Sixth Affiliated hospital, School of Life Sciences, Sun Yat-sen University, Guangzhou 510060, China

²Guangdong Provincial Key Laboratory of Pharmaceutical Bioactive Substances, School of Biosciences and Biopharmaceutics, Guangdong Pharmaceutical University, Guangzhou 510006, China

³State Key Laboratory of Oncology in South China, Cancer Center, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University, Guangzhou 510060, China

⁴Institute of Precision Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510060, China

⁵Guangdong Institute of Gastroenterology, Biomedical Innovation Center, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510060, China

*Correspondence address. Yubin Xie, Institute of Precision Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510060, China.

E-mail: xieyb6@mail.sysu.edu.cn; Xiaotong Luo, Innovation Center of the Sixth Affiliated hospital, School of Life Sciences, Sun Yat-sen University, Guangzhou 510060, China. E-mail: luoxt36@mail.sysu.edu.cn

[†]Yin Zhang, Lin Tang, and Shengyao Zhi contributed equally to this work.

Abstract

Background: Allelic gene-specific regulatory events are crucial mechanisms in organisms, pivotal to many fundamental biological processes such as embryonic development and chromosome inactivation. Allelic gene imbalance manifests at both RNA expression and epigenetic levels. Recent research has unveiled allelic-specific regulation of RNA N⁶-methyladenosine (m⁶A), emphasizing the need for its precise identification. However, prevailing approaches primarily focus on screening allele-specific genetic variations associated with m⁶A, but not truly identify allelic m⁶A events. Therefore, the construction of a novel algorithm dedicated to identifying allele-specific m⁶A (ASm⁶A) signals is still necessary for comprehensively understanding the regulatory mechanism of ASm⁶A.

Findings: To address this limitation, we have developed a meta-analysis approach using hierarchical Bayesian models to accurately detect ASm⁶A events at the peak level from MeRIP-seq data. For user convenience, we introduce a unified analysis pipeline named M6Allele, streamlining the assessment of significant ASm⁶A across single and paired samples. Applying M6Allele to MeRIP-seq data analysis of pulmonary fibrosis and lung adenocarcinoma reveals enrichment of ASm⁶A events in key regulatory genes associated with these diseases, suggesting their potential involvement in disease regulation.

Conclusions: Our effort provides a method for precisely identifying ASm⁶A events at the peak level, elucidates the interplay of m⁶A with human health and disease genetics, and paves a new visual angle for disease research. The M6Allele software is freely available at <https://github.com/RenLabBioinformatics/M6Allele> under the MIT license.

Keywords: allele-specific, RNA N⁶-methyladenosine (m⁶A), hierarchical Bayesian model, meta-analysis

Introduction

In a non-haploid genome, the transcriptional activity at different gene alleles can vary significantly [1]. Allele-specific effects are crucial in various cellular activities, particularly genomic imprinting [2], chromosome inactivation [3], and the regulation of gene expression in particular spatiotemporal circumstances [4]. Mechanisms such as random monoallelic expression [5, 6], allele sequence-specific expression, and parental-specific (imprinted) expression [7, 8] have been shown to result in the expression of only one allele for many genes. Allele-specific gene expression (ASE) can impact disease traits, including biological developmental abnormalities [9], cardiovascular and cerebrovascular dysfunctions [10], progressive genetic disorders [11], and even cancers [12, 13]. In addition to ASE, allelic imbalance is evident in epigenetic regulation. Extensive research has focused on allele-specific DNA methylation as a factor that controls allele-specific expression [14, 15]. Notably, approximately 10% of human genes are regulated by allele-specific DNA methylation [16]. While these studies primarily focused on DNA-level modifications that influ-

ence allele-specific regulation, RNA-level modifications have received less attention.

Similar to DNA methylation, RNA methylation is a common and reversible epigenetic modification found in RNA nucleotides. Among all the RNA methylation types, N⁶-methyladenosine (m⁶A) is the most common modification in eukaryotic messenger RNAs (mRNAs), accounting for over 80% of known RNA modifications [17]. m⁶A is also extensively present in microRNAs (miRNAs) [18], long noncoding RNAs (lncRNAs) [19], and circular RNAs (circRNAs) [20]. m⁶A is widely involved in a variety of important cell processes, including embryonic development [21], apoptosis [22], and sperm development [23], as well as in a large number of malignant diseases, such as tumors and obesity [24, 25]. Therefore, m⁶A is a key factor for understanding disease pathogenesis and developing new therapies.

Recent studies have revealed that allelic regulations also existed in m⁶A modifications [26]. For example, Olazagoitia-Garmendia et al. have shown that allele-specific m⁶A (ASm⁶A) in lncRNAs, such as LOC339803, affects protein binding and

Received: September 4, 2024. Revised: January 5, 2025. Accepted: March 7, 2025

© The Author(s) 2025. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

chromatin localization, and that an SNP in the 5'UTR of XPO1 associated with celiac disease, which is close to three m⁶A consensus motifs (GGACT), exhibits higher m⁶A methylation, leading to increased XPO1 protein levels and activation of nuclear factor kappa B (NFkB), contributing to inflammation [27, 28]. To identify the transcriptome-wide ASm⁶A, Cao et al. recently applied Fisher's exact test to detect ASm⁶A at the SNP level in MeRIP-seq data. They identified 12,056 allele-specific SNPs located in m⁶A peaks from human tissues and found that many of them are associated with risk variants in common diseases [29]. In addition, Yi et al. developed ASPRIN [30] (Allele-Specific Protein–RNA Interaction) to identify genetic variations that alter RBP–RNA interactions by jointly analyzing CLIP-seq and RNA-seq data, which can theoretically also be applied to analyze variant sites associated with ASm⁶A on MeRIP-seq. However, both of these methods only estimated the allele-specific imbalance of m⁶A peaks at the SNP level, rather than truly identifying allelic m⁶A events, making it challenging to interpret the underlying mechanisms of ASm⁶A in different biological processes. Since MeRIP-seq provides modification peaks of approximately 200 nt, actual data demonstrate that a significant number of m⁶A peaks possess multiple detectable heterozygous SNPs. This underscores the importance of having a framework for integrating expression information across individual sites in a peak region to accurately assess allele-specific imbalance of m⁶A. However, there is currently no standard or robust method for summarizing information across SNPs into a single measure of ASm⁶A for the entire peak.

To overcome these difficulties, we developed a new ASm⁶A detection method, named M6Allele, which uses a hierarchical Bayesian model to assess ASm⁶A by integrating information across individual heterozygous SNPs within a peak, even without any prior knowledge of haplotype phasing [31]. Our approach demonstrates higher precision and fewer false positives compared with previous methods using Fisher's exact test. For users' convenience, we have built a comprehensive toolkit for the one-stop analysis of ASm⁶A from MeRIP-seq data [32]. We applied M6Allele to a panel of human pulmonary fibrosis tissues and paired tumor–normal lung tissue samples. The results indicated enrichment of disease-specific ASm⁶A modifications in pathogenic genes, suggesting a potential role for ASm⁶A in disease regulation. Our study introduced a novel meta-analytic approach that enables the precise and sensitive analysis of the dynamic network of ASm⁶A at the peak level. This method facilitates the identification of specific m⁶A modifications occurring at the allele level, as well as the comprehension of their association with human health and disease.

Results

M6Allele: meta-analysis based detection of ASm⁶A modifications

In this study, we introduce M6Allele, an algorithm designed for identifying ASm⁶As in MeRIP-seq data (Fig. 1A). Initially, high-confidence heterozygous SNVs were identified through variant calling, with rigorous filtering applied to mitigate transcription and mapping biases [33]. Variants were retained if they were absent in RNA editing sites (RADAR database) [34] but present in the dbSNP database. Subsequently, we calculated the read counts of alleles from m⁶A IP and input sequencing data, followed by a hierarchical Bayesian model to evaluate the modification difference between the two alleles at individual SNPs within a modification unit. For M6Allele, we chose peaks as units, which can be obtained through peak calling tools commonly used in MeRIP-seq

data analysis. Therefore, we only considered SNPs located in the peak regions.

M6Allele requires prior knowledge of gene haplotype specifications, which are likely unknown for the MeRIP-seq dataset. To determine the allelic origin specificity of reads, we adopted MBASED's strategy [35] and introduced a pseudo-phasing approach for SNPs. Specifically, for each retained SNP, we counted the frequencies of different base types in the m⁶A Input sample separately. The two base types with the highest frequencies are assigned as the “major” and “minor” haplotypes, respectively. To precisely detect allelic imbalance within peaks, we quantified it as the odds ratio of the major allele relative to the minor allele in the m⁶A IP sample. The detection in ASm⁶A then became the identification of peaks with an odds ratio significantly > 1.

To accurately evaluate the allelic imbalance of a m⁶A peak, we utilized a random-effects model (REM) [36] to integrate the odd ratios of all SNPs within the peak. Essentially, we considered the regression coefficients of the fixed-effects model (in our case, ASm⁶A) as random variables, assuming all coefficients follow the same normal distribution. By constructing a hierarchical Bayesian model, we estimated the mean of this normal distribution, which served as the ASm⁶A odds ratio for the entire peak. Similarly, to remove the influence of ASE on assessing ASm⁶A events, we constructed a hierarchical Bayesian model for ASE using m⁶A Input samples at the gene level. The odds ratio of ASE obtained served as the background odds ratio for the ASm⁶A model.

Because of the pseudo-phasing strategy used by M6Allele to infer gene haplotypes, the statistical significance of ASm⁶A may lead to anticonservative nominal P-values [36]. To effectively address this issue, we simulated MeRIP-seq data using SNP loci from the 1000 Genomes Project [37] and dbSNP [38] databases to mimic the absence of allele-specific events. We introduced the generalized Pareto distribution (GPD) [39] for fitting the deviation of allelic odds ratio under pseudo-phasing, to adjust the statistical significance level. M6Allele converts the odds ratio of each peak into the frequency of the major allele (MAF) and provides its corresponding P-value. By adjusting the P-values using the Benjamini–Hochberg (BH) method [40], we obtain Q-values. A peak with a Q-value below 0.05 is considered a significant allelic m⁶A imbalance event.

Additionally, M6Allele includes a paired-sample analysis module for detecting differential ASm⁶A between paired samples from the same individual. Given that true haplotypes are unknown, maintaining the consistency of haplotypes across paired samples involves designating one as the control group of pseudo-phasing. For instance, in a tumor versus normal comparison, we designate the normal sample as the control group and classify haplotypes into “major” and “minor” alleles based on read counts obtained from the Input sample. Differences between m⁶A odds ratio at an individual SNP in the two samples are used as measures of sample-specific ASm⁶A. SNP-level scores are combined into a peak-level score using meta-analysis and a hierarchical Bayesian model, which is analogous to the single-sample approach. This composite odds ratio provides an estimate of the peak-level odds ratio difference between samples.

The details of M6Allele are provided in Methods and Supplementary Methods. Our algorithm is implemented in Java, and the corresponding JAR file has been built. For users' convenience, we have developed an integrated pipeline for ASm⁶A analysis using Docker (<https://www.docker.com/>) (Fig. 1B).

Robust ASm⁶A detection by M6Allele

Because of the absence of gold-standard MeRIP-seq data featuring allele-specific events, we aimed to evaluate the performance of

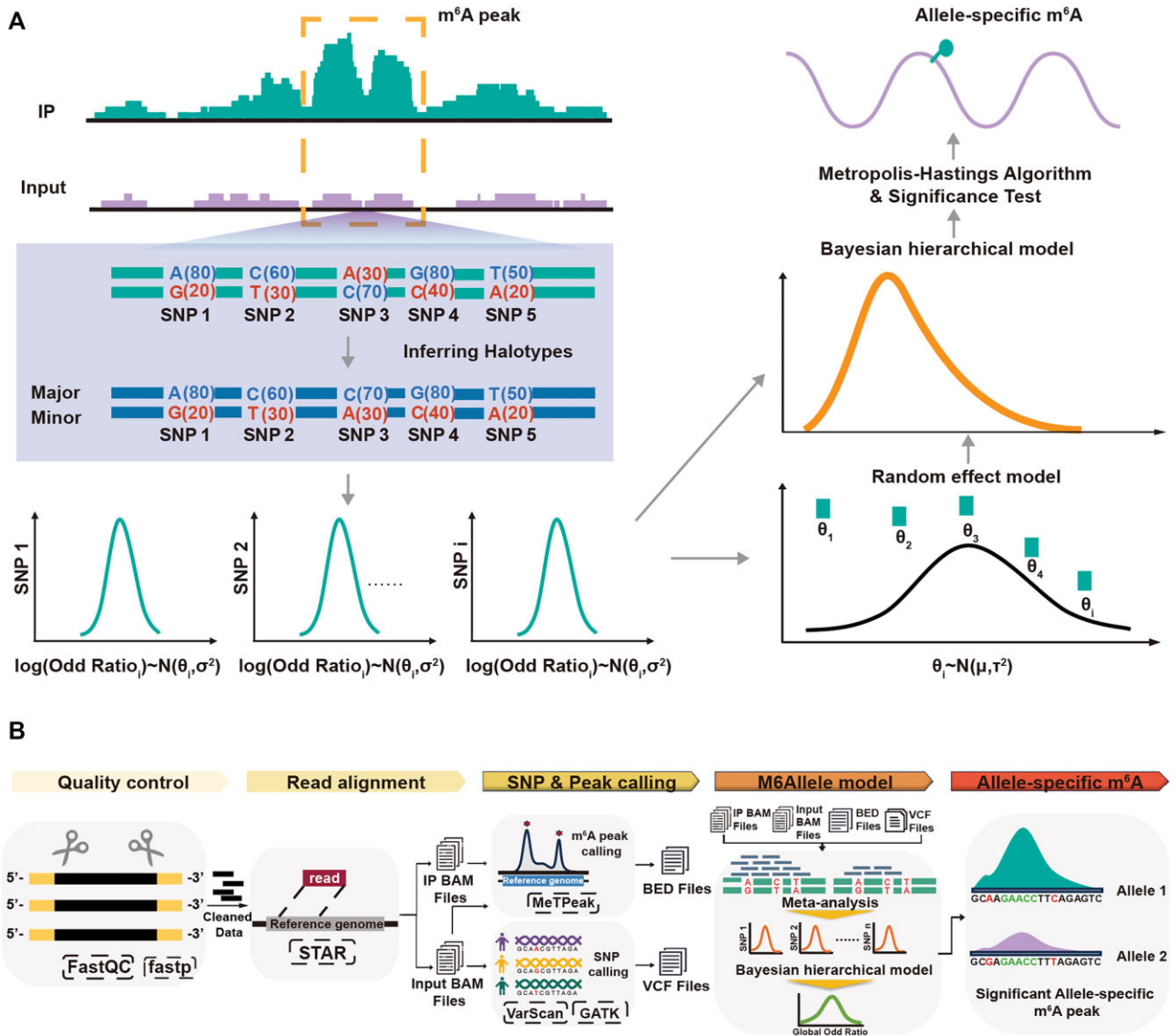


Figure 1: ASm⁶As analysis pipeline. (A) Schematic diagram of the M6Allele model. (B) ASm⁶As identification pipeline based on Docker.

M6Allele in the absence of phasing information using simulated MeRIP-seq data. The simulation process detailed in [Additional file 1: Fig. S1](#), draws inspiration from the methods of Polyester [41] for simulating RNA-seq.

Because accurate ASE results are essential for M6Allele to assess ASm⁶A effectively, we initially evaluated the ASE detection performance of M6Allele using simulated RNA-seq data. During the ASE simulation, 50% of transcripts were randomly selected to represent positive ASE events. For these transcripts, the MAF was uniformly sampled from [0.6, 0.9], while the rest had an MAF of 0.5. Additionally, to assess M6Allele's robustness in identifying significant ASE events, we simulated RNA-seq data with different sequencing read lengths (75, 100, 150, and 300 nt) 50 times each. Then, we applied M6Allele's ASE detection method to each simulated dataset, considering genes with a Q-value ≤ 0.05 as significant ASE events. Among the current ASE detection tools, GeneiASE [42] and MBASED [35] can only utilize RNA-seq data to identify ASE events. Consequently, we conducted a performance comparison of M6Allele with these tools (Fig. 2A–E). We observed that the overall precision of M6Allele remains robust across various simulated

sequencing read lengths, showing minimal impact (Fig. 2A). However, recall increases with longer read lengths (Fig. 2B). We maintained the overall false discovery rate (FDR) at a nominal level of 5%, affirming the effectiveness of P-value adjustment (Fig. 2C). By integrating precision and recall results, we calculated the F0.5 and F1 scores [43] to comprehensively assess the performance of M6Allele in ASE identification (Fig. 2D,E). Comparing M6Allele to two other ASE detection tools reveals its consistently superior performance ([Additional file 2: Table S1](#)), indicating its precision in ASE detection is suitable for downstream analysis. To further validate M6Allele's ASE detection performance on real data, we used M6Allele to identify ASE in the RNA-seq data GSM4998283. Among the results, we chose a gene (RMRP) with significant ASE and one (H1-3) without significant ASE. Visualization with the IGV tool (Fig. 2F) showed their haplotype distributions, confirming M6Allele's accurate identification of ASE events, consistent with reality.

We subsequently assessed the detection performance of ASm⁶A by M6Allele using simulated data. To ensure the simulated dataset accurately reflected the genuine peak lengths

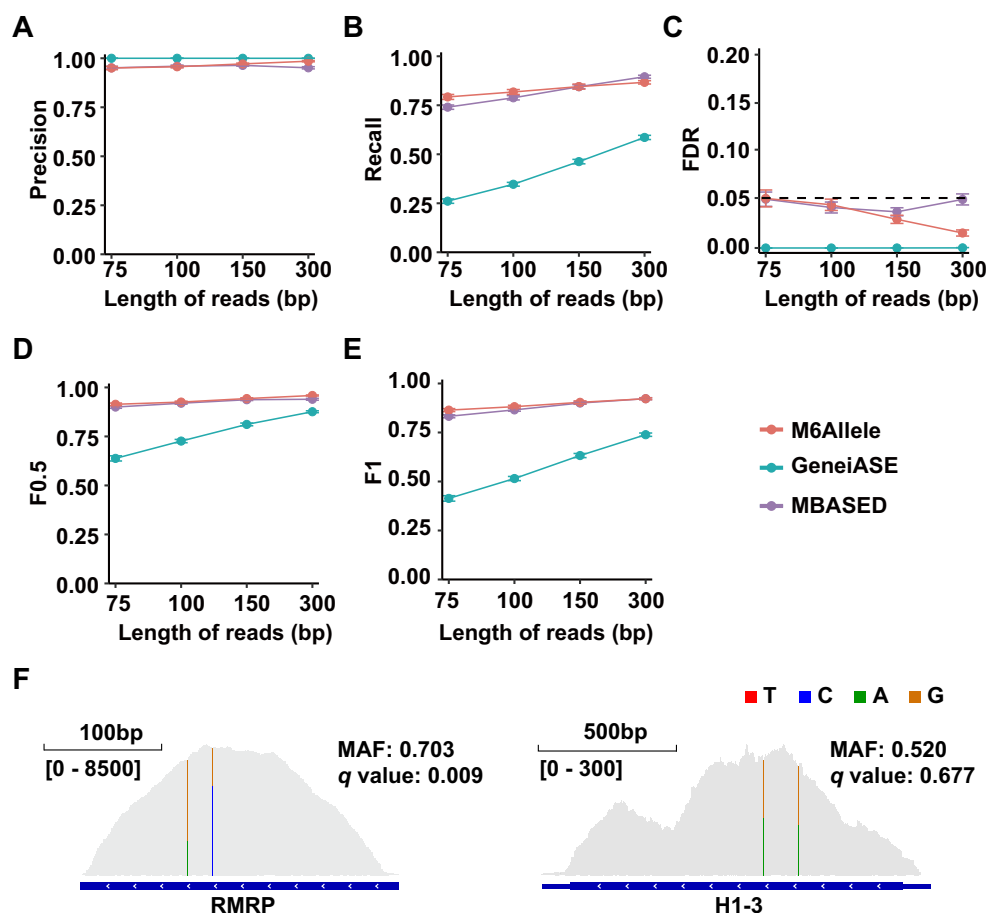


Figure 2: Performance comparison of different tools for ASE identification on simulated datasets. (A) The precision rates in ASE analysis between M6Allele, MBASED, and GeneiASE with different numbers of biological replicates. (B) The recall rates in ASE analysis between M6Allele, MBASED, and GeneiASE with different numbers of biological replicates. (C) The false discovery rates in ASE analysis between M6Allele, MBASED, and GeneiASE with different numbers of biological replicates. (D) The F0.5 scores in ASE analysis between M6Allele, MBASED, and GeneiASE with different numbers of biological replicates. (E) The F1 scores in ASE analysis between M6Allele, MBASED, and GeneiASE with different numbers of biological replicates. (F) Visualization of the number of reads covered by allele-specific expressed gene versus non-allele-specific expressed gene.

and distribution of m^6A modifications, we incorporated ASm^{6A} events into the simulation by leveraging m^6A peaks and sites from GSM1828594. Moreover, for a comprehensive analysis of M6Allele's performance, we categorized all test peaks within the samples based on five pertinent variables: read lengths, library size, FPKM of gene expression, the number of SNPs in a peak, and the number of biological replicates. In each category, 50% of the peaks were randomly designated as allele-specific, i.e., true positives for ASm^{6A} (MAF > 0.6), while the rest were labeled as true negatives for ASm^{6A} (MAF = 0.5). For robust evaluation, each simulated dataset was repeatedly analyzed 50 times. The results demonstrated that changes in sequencing read length do not affect the performance of M6Allele (Fig. 3A). However, increases in library size, FPKM, the number of SNPs in a peak, and the number of biological replicates led to a reduction in the average error rate, with particularly pronounced improvements observed for greater library depth and higher gene expression levels (Fig. 3B–E). Despite these variations, in simulated data tests, M6Allele consistently maintained an error rate below 10%, even in small libraries or for genes with low expression levels. This underscores the robustness of the M6Allele model and demonstrates its applicability to sequencing data across diverse experimental conditions.

Furthermore, we compared the performance of M6Allele with two additional tools capable of detecting ASm^{6A} events, ASPRIN

[30] and the algorithm developed by Cao et al. [29]. We followed the tutorials provided by the two tools, sticking to their default parameter settings. As these tools can only obtain individual SNP sites associated with ASm^{6A}, we aligned the SNPs with m^6A peaks. If any SNP within a peak was identified as having ASm^{6A} modification by ASPRIN or Cao et al.'s algorithm, the peak was classified as ASm^{6A} modified, resulting in a positive outcome; otherwise, it was considered non-ASm^{6A}. According to the ASm^{6A} detection results from various algorithms, we calculated the area under the ROC curve (AUC) for each category of simulated peaks. The results indicated that, across diverse settings of the simulated data, M6Allele consistently exhibits a significantly higher average AUC compared with the other two algorithms (Fig. 3F; [Additional file 1: Fig. S2](#)). Furthermore, we also investigated the impact of the pseudo-phasing strategy on the performance of M6Allele. The results indicated that the AUC of M6Allele was 0.9216 with known phasing information, which is comparable to the AUC of 0.9039 obtained with pseudo-phasing strategy (Fig. 3F). This further validates the reliability of the pseudo-phasing method in ASm^{6A} detection. Since the other two methods identify SNP sites related to ASm^{6A}, they were more susceptible to the influence of different sequencing conditions. As the observed data on peaks increased, such as the number of covered SNPs or biological replicates, their performance improved significantly. This suggested

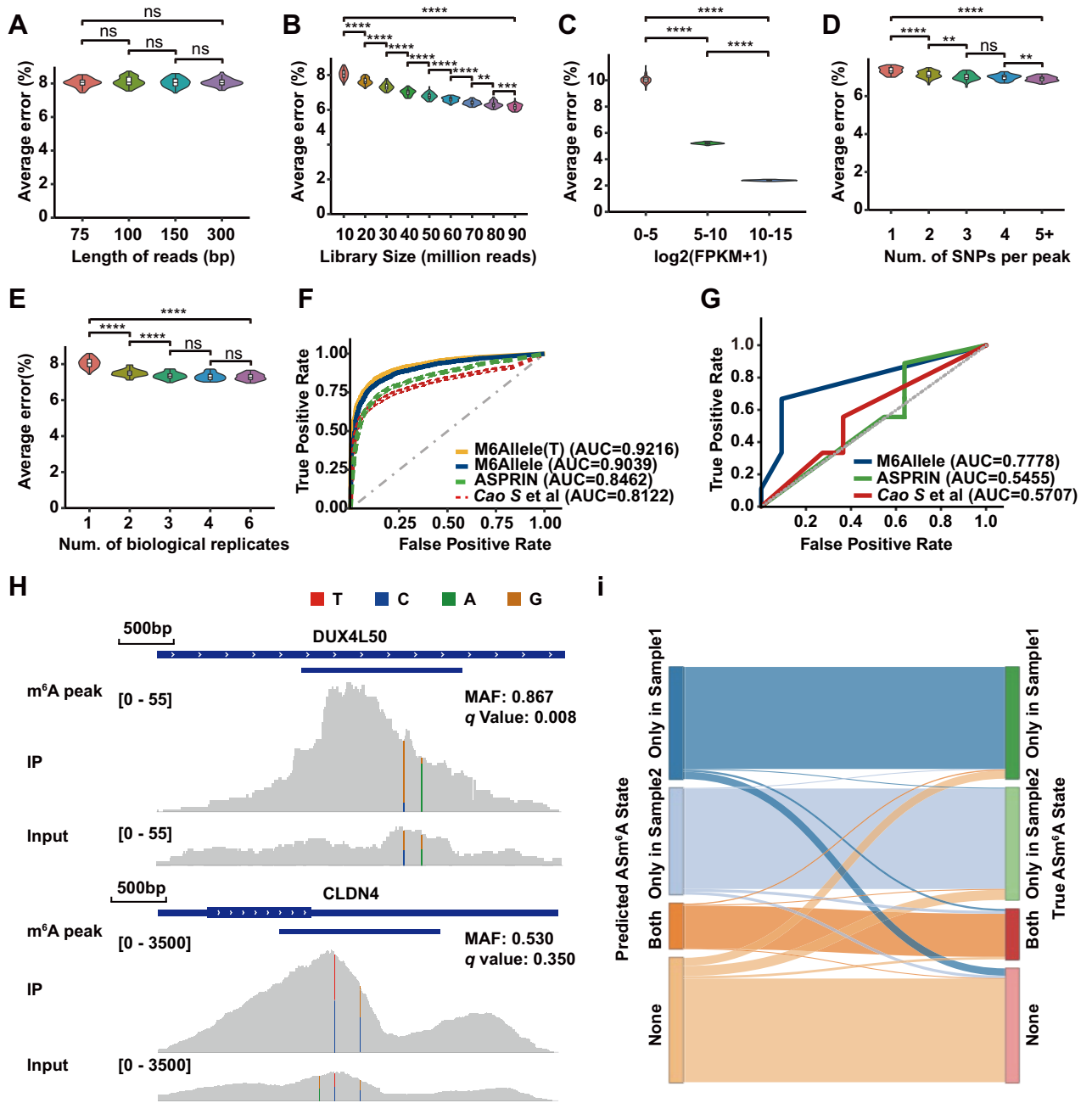


Figure 3: Evaluation of the M6Allele algorithm performance on various metrics. (A) The error in ASm⁶A analysis in M6Allele with different sequencing lengths. (B) The error in ASm⁶A analysis in M6Allele with different library sizes. (C) The error in ASm⁶A analysis in M6Allele with different FPKMs. (D) The error in ASm⁶A analysis in M6Allele with different numbers of SNP sites covered by each modification peak. (E) The error in ASm⁶A analysis in M6Allele with different numbers of biological replicates. (F) Performance evaluation and comparison of M6Allele, ASPRIN, and the algorithm developed by Cao et al. in the simulated MeRIP-seq dataset. (G) Performance evaluation and comparison of M6Allele, ASPRIN, and the algorithm developed by Cao et al. using the MeRIP-seq data from THP-1 cell line, followed by Sanger sequencing. (H) Visualization of the number of reads covered by ASm⁶A peak versus non-ASm⁶A peak. (I) Identification of sample-specific ASm⁶A events in the simulated paired-samples dataset.

that the approach of relying solely on individual SNPs to identify ASm⁶A may struggle to avoid errors caused by the noise of sequencing data. To provide a more comprehensive evaluation, we further compared the detailed performance metrics of different ASm⁶A detection algorithms (Additional file 1: Fig. S3). The results showed that M6Allele consistently outperformed the other two tools across all categories, combining higher precision and recall while maintaining a lower false discovery rate. Notably, M6Allele exhibited smaller fluctuations and superior stability compared with the other two tools, especially under challenging conditions

such as lower library sizes or fewer biological replicates. This underscores the robustness of M6Allele, particularly compared to Cao et al.'s algorithm, which showed significant performance improvement with increasing gene expression, potentially indicating its relatively higher restriction on the number of reads and lower sensitivity in identifying ASm⁶A signals in low-expressed genes.

Additionally, we evaluated the computational time of M6Allele in comparison to two other tools for both ASE and ASm⁶A detection tasks. Using two publicly available MeRIP-seq datasets from GEO database (GSE164151 and GSE198288) with a total of 12

human samples, we tested computational efficiency across five sample size gradients under single-threaded mode. The results showed that M6Allele exhibited comparable speed to geneiASE for ASE detection and intermediate performance for ASm⁶A detection, being slower than Cao et al.'s tool but faster than AS-PRIN (Additional file 1: Fig. S4). This difference in speed may be attributed to M6Allele's more comprehensive integration of SNP information within peaks, which increases computational complexity while ensuring higher detection accuracy.

To further validate M6Allele's ASm⁶A detection performance, we conducted experimental validations using MeRIP-seq (GSE289760) on the human monocytic THP-1 cell line, followed by Sanger sequencing. A total of 20 candidate sites were selected based on predictions from M6Allele (Additional file 2: Table S2, Table S3, Table S4). We utilized EditR software to analyze the Sanger sequencing chromatograms, calculating the ratios of different nucleotides at the selected sites in both the IP and input samples, and determining the odds ratio for the major allele. Using a threshold of greater than 1.2 for the odds ratio, we classified 9 ASm⁶A sites as true positives and 11 as true negatives. Using these 20 sites, we compared the performance of three ASm⁶A detection tools. The results demonstrated that M6Allele achieved a significantly higher AUC compared to the other two tools (Fig. 3G). Additionally, we analyzed the IP and Input samples from GSE164151 (GSM4998285 and GSM4998284). From the results, we randomly selected three peaks with significant ASm⁶A and three peaks showing no significant ASm⁶A for visualization using IGV (Fig. 3H and Additional file 1: Fig. S5). Their haplotype distributions in IP and Input samples confirmed M6Allele's precise identification of ASm⁶A events, aligning with actual observations.

Similarly, simulations were performed in the paired-sample setting (Supplementary Methods). To evaluate M6Allele's accuracy for detecting sample-specific ASm⁶A events in the paired-sample analysis, MeRIP-seq data for paired-samples were generated using identical genotypic and m⁶A peaks. Then, 956 peaks were randomly classified into four ASm⁶A categories: absent in both samples, present only in sample 1, present only in sample 2, and present in both samples. Through paired-sample analysis using M6Allele and comparing the results with the peak assignments (Fig. 3I), precise identification of sample-specific ASm⁶A events was observed, achieving an overall accuracy of 89.9%. To illustrate these four ASm⁶A categories, we provided IGV visualizations of randomly selected examples for each category (Additional file 1: Fig. S6). The observed haplotype distributions in IP and Input samples were consistent with M6Allele's detection results for differential ASm⁶A events.

ASm⁶A modifications are closely associated with pulmonary fibrosis

The impact of ASm⁶A modification on human diseases is our focal point. However, only a few studies report an association between ASm⁶A and diseases. Previous studies demonstrated that pulmonary fibrosis is a typical disease regulated by m⁶A modification. To further investigate the impact of ASm⁶A modification on pulmonary fibrosis, we utilized M6Allele to analyze the distribution of ASE and ASm⁶A events in patients with pulmonary fibrosis (Additional file 2: Table S5, Table S6). We identified widespread ASE and ASm⁶A modifications across 22 pairs of autosomal chromosomes in patients with pulmonary fibrosis (Fig. 4A,B). Compared to normal human tissue, we found 111 genes exhibiting significant ASE exclusively in all pulmonary fibrosis patient tissues (re-

ferred to as ASE-Gain), along with 94 genes showing significant ASE only in normal tissue (referred to as ASE-Loss) at the whole-genome level (Additional file 1: Fig. S7a,b). Similarly, we detected 64 specific ASm⁶A-modified genes (ASm⁶A-Gain) and 62 genes with ASm⁶A-Loss in pulmonary fibrosis patient tissues. We found very few genes shared between ASE and ASm⁶A, with only six genes showing a gain of both ASE and ASm⁶A, and just one gene showing a loss of both ASE and ASm⁶A (Additional file 1: Fig. S7a, b). This suggested that ASm⁶A may exert its regulatory function through alternative mechanisms instead of only impact allelic gene expression. Next, we conducted pathway enrichment analysis on genes associated with ASE and ASm⁶A events with a FDR < 0.05, utilizing the "GO Biological Processes" dataset from the Metascape database [44] (Fig. 4C,D and Additional file 1: Fig. S7c,d). In the patient tissues, genes with ASm⁶A-Gain were significantly enriched in immune response, complement activation classical pathway, Rho protein signaling, and other functional pathways closely related to human pulmonary fibrosis disease (Fig. 4C). Meanwhile, genes exhibiting ASE-Gain in pulmonary fibrosis were enriched in aorta morphogenesis, response to interferon-gamma, negative regulation of cell growth, and other pathways related to lung vasculature, cell, and immunity (Additional file 1: Fig. S7c). ASm⁶A-Loss genes were visibly enriched in pathways associated with epithelial cell differentiation, MAP kinase activation, changes in cell morphology, immune activation response, and platelet-derived growth factors associated with pulmonary fibrosis diseases (Fig. 4D). Genes with ASE-Loss in pulmonary fibrosis played crucial roles in growth factor and metabolism-related pathways (Additional file 1: Fig. S7d). These results suggested that ASE and ASm⁶A events may collectively influence the development of pulmonary fibrosis through interconnected pathways.

To further elucidate the regulatory relationship between ASm⁶A and pulmonary fibrosis, we compared known pulmonary fibrosis-related genes (Score_{GDA} ≥ 0.3) from the DisGeNET database [45] with ASm⁶A-Gain and ASm⁶A-Loss genes in pulmonary fibrosis patients. Initially, we conducted a hypergeometric test to analyze the relationship between ASm⁶A-Gain and ASm⁶A-Loss genes and known pulmonary fibrosis genes, utilizing all annotated genes in the GTF file of hg38 as a sample population, totaling 58,676 genes. The result ($P < 5 \times 10^{-7}$) revealed a significant enrichment of ASm⁶A-modified genes within the pulmonary fibrosis gene set (the deep blue bar in Fig. 4E). To delve deeper into the regulation of pulmonary-fibrosis-associated genes by ASm⁶A modification, we identified genes interacting with ASm⁶A-modified genes with confidence of 0.9 from the STRING database [46] and determined their overlap with pulmonary-fibrosis-related genes. In ASm⁶A-modified genes and their interactors, referred to as ASm⁶A-regulated genes, hypergeometric testing unveiled a significant enrichment of pulmonary-fibrosis-related genes (the green bar in Fig. 4E). Functional pathway analysis of this gene overlaps highlighted significant enrichment in pathways crucial to pulmonary fibrosis pathogenesis, notably positive regulation of phosphorylation [47], negative regulation of cell differentiation [48], and positive regulation of immune response [48] (Additional file 1: Fig. S8a). Additionally, we conducted a similar analysis on ASE genes (Additional file 1: Fig. S8b). The overlapping genes showed enrichment in pathways such as positive regulation of cell migration [49], response to growth factor [50], and negative regulation of cell differentiation [51]. The hypergeometric test between ASE genes and pulmonary fibrosis-related genes revealed a significant enrichment of ASE genes among pulmonary-fibrosis-related genes (the orange bar in Fig. 4E). Meanwhile, pulmonary-fibrosis-related genes were also significantly enriched among ASE genes

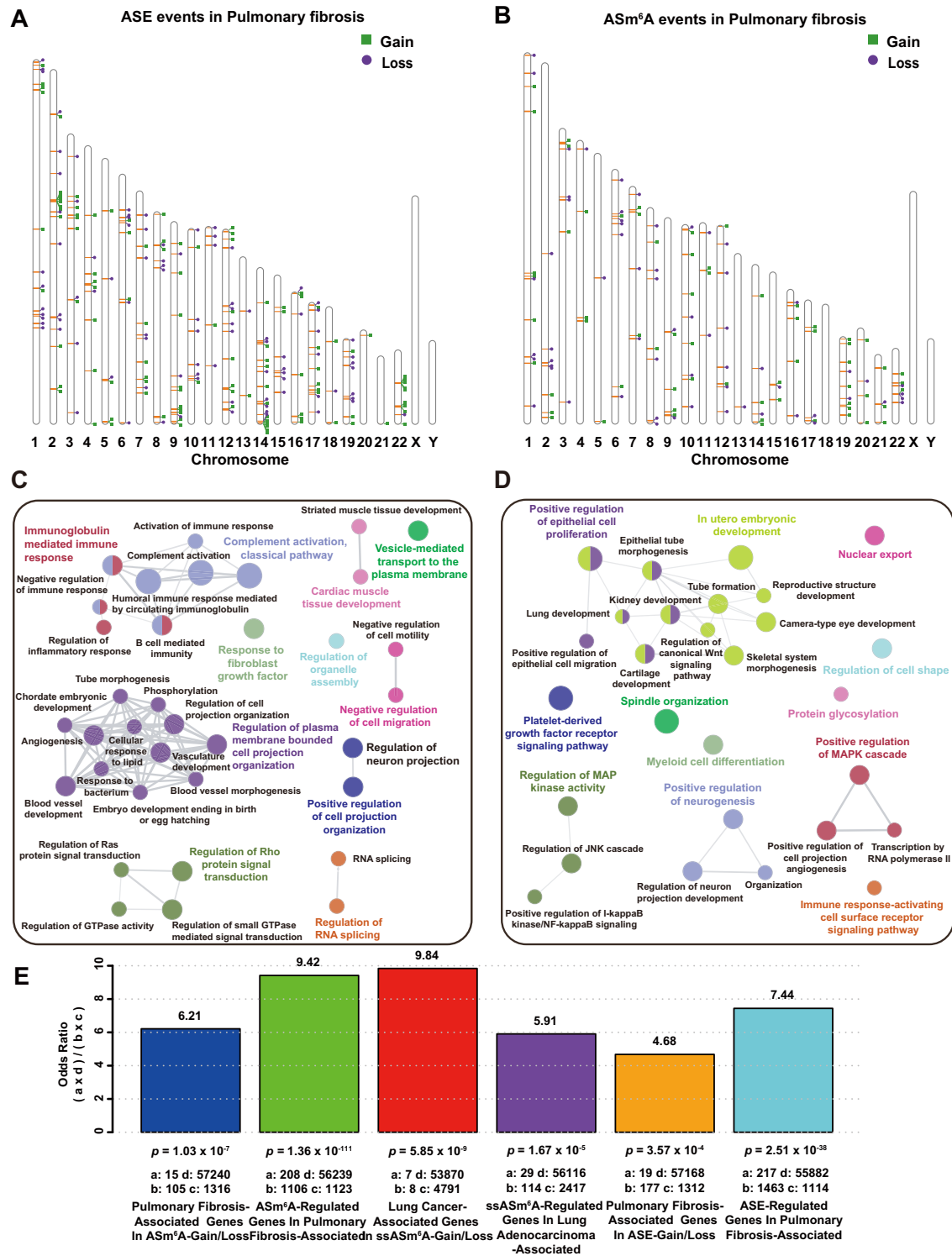


Figure 4: Analysis results of M6Allele on pulmonary fibrosis dataset. (A) The chromosomal distribution of genes with ASE. (B) The chromosomal distribution of genes with ASm⁶A. (C) The results of GO enrichment analysis with ASm⁶A-Gain genes. (D) The results of GO enrichment analysis with ASm⁶A-Loss genes. (E) The hypergeometric test results for ASm⁶A-Gain/Loss and ASE-Gain/Loss events related to disease-associated genes in pulmonary fibrosis and lung adenocarcinoma. The deep blue bar represents the odds ratio of genes associated with pulmonary fibrosis observed in ASm⁶A-Gain/Loss genes, where “a” denotes the overlap between ASm⁶A-Gain/Loss genes and pulmonary fibrosis-associated genes, “b” represents ASm⁶A-Gain/Loss genes exclusively, “c” denotes genes exclusively associated with pulmonary fibrosis, and “d” represents genes that do not belong to either category. The red bar represents the odds ratio of genes associated with lung cancer observed in ssASm⁶A-Gain/Loss genes of tumor samples. The green bar represents the odds ratio of ASm⁶A-regulated genes observed in pulmonary fibrosis-associated genes, where “a” denotes the overlap between both gene categories, “b” represents genes exclusively associated with pulmonary fibrosis, “c” denotes genes exclusively regulated by ASm⁶A, and “d” represents genes that do not belong to either category. The meanings of “a” to “d” in the remaining bars are analogous to those in above two bars. The purple bar represents the odds ratio of genes associated with lung adenocarcinoma observed in ssASm⁶A-regulated genes of tumor samples. The orange bar represents the odds ratio of genes associated with pulmonary fibrosis observed in ASE-Gain/Loss genes, and the light blue bar represents the odds ratio of ASE-regulated genes observed in pulmonary fibrosis-associated genes.

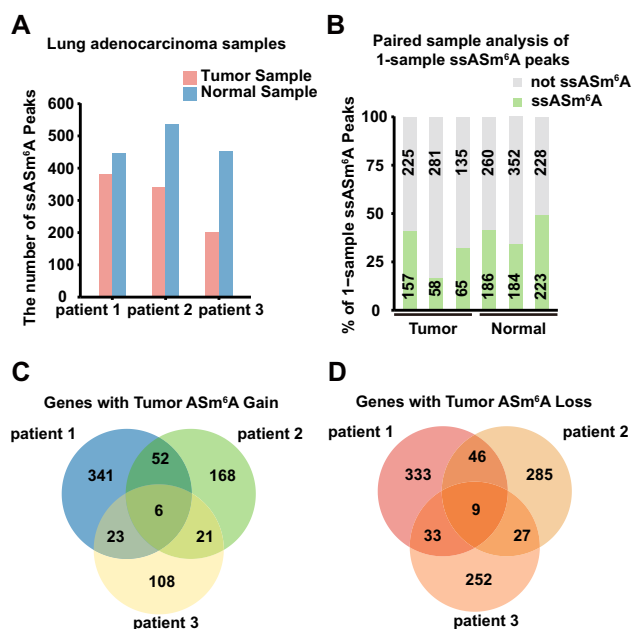


Figure 5: Analysis results of M6Allele on lung adenocarcinoma dataset. (A) ssASm⁶A events identified by the single-sample analysis strategy. (B) Comparison of the ssASm⁶A events from the single-sample analysis and the paired-sample analysis. (C) ssASm⁶A-modified genes in tumor samples. (D) ssASm⁶A-modified genes in normal samples.

and their interactors (the light blue bar in Fig. 4E). The above findings suggest that genes with allele-specific events identified by M6Allele may interact with known pulmonary-fibrosis-related genes, regulate related pathways, and thus influence the progression of pulmonary fibrosis diseases. M6Allele can unearth ASm⁶A-modified genes closely related to diseases from the MeRIP-seq data, providing a new direction for research on the pathogenesis and treatment of human diseases.

M6Allele reveals lung-adenocarcinoma-associated ASm⁶A with high heterogeneity

It has previously been reported that m⁶A modification can regulate the occurrence and development of cancers [52], particularly lung adenocarcinoma, a malignant tumor with an exceptionally high mortality rate [53, 54]. Notably, there have been no reports on whether ASm⁶A modification regulates the progression of malignant tumors. To further explore the impact of ASm⁶A modification on lung adenocarcinoma, we used M6Allele to identify ASm⁶A events in lung adenocarcinoma patients [55] (Additional file 2: Table S7). In cancer research, we typically emphasize intergroup differences between tumors and adjacent tissues unaffected by individual genetic information, such as sample-specific ASm⁶A (ssASm⁶A) events. As ASm⁶A events achieved from unpaired-sample analysis of tumor samples often include many events unrelated to the disease, such as the patient's inherited ASm⁶A events, filtering out these false positives is crucial for identifying disease-relevant ASm⁶A modifications. Therefore, we compared two strategies, the unpaired-sample and paired-sample analysis, to exclude false-positive ssASm⁶A events. The results of the single-sample analysis showed that tumor samples from three patients had 382, 339, and 651 peaks with ASm⁶A, while in the normal samples, there were 446, 536, and 451 peaks with ASm⁶A (Fig. 5A). Through paired-sample analysis, we found that only 17–49% of the ASm⁶A events identified in unpaired-sample analysis

were recognized as single-sample ASm⁶A signals (Fig. 5B). The remaining ASm⁶A signals were present in both tumor and normal samples, suggesting these events may be inherent epigenetic regulatory events in patients unrelated to the tumor. These results illustrate that paired-sample analysis can effectively screen for ssASm⁶A modifications and identify significant differences in ASm⁶A events between samples. Therefore, in downstream analysis, we focused solely on single-sample ASm⁶A events.

To examine the uniformity of ssASm⁶A sites among different patient samples, we combined the analysis results to create a Venn diagram. The results reveal that, among the 422, 247, and 158 tumor ssASm⁶A-Gain genes identified in the three patient samples, only six genes were shared (Fig. 5C). Similarly, there were only nine shared tumor ssASm⁶A-Loss genes in the three patient samples, while the identified genes numbered 421, 367, and 321, respectively (Fig. 5D). These findings indicate that the tumor ssASm⁶A-modified genes identified in different patients with lung adenocarcinoma differ significantly, and the gain and loss of ASm⁶A also vary notably across different patient samples. Moreover, the proportion of identified ASm⁶A-modified genes existing alone in a single sample accounted for as much as 79.95% (674 out of 843), 73.78% (453 out of 614), and 75.16% (360 out of 479), respectively. These results highlight the highly heterogeneous and complex nature of lung-adenocarcinoma-associated ASm⁶A modifications.

To prove the effectiveness of the algorithm, we annotated the 15 ssASm⁶A genes shared among the three patient samples through a literature review (Additional file 2: Table S8). Among these, five genes were reported to be directly associated with lung adenocarcinoma, four genes were associated with lung cancer but not specifically with lung adenocarcinoma, and six genes were not reported to be related to lung cancer but were found to be associated with other types of cancer. Furthermore, our hypergeometric testing revealed significant enrichment of these 15 genes among lung-cancer-related genes from DisGeNet (Score_{GDA} ≥ 0.3; the red bar in Fig. 4E). These findings demonstrate the algorithm's effectiveness in identifying significant genes related to cancer. In addition, genes associated with lung adenocarcinoma (Score_{GDA} ≥ 0.3 in DisGeNet) were significantly enriched among ssASm⁶A-regulated genes including ssASm⁶A-modified genes and their interactors (confidence ≥ 0.9 in STRING) (the purple bar in Fig. 4E). This suggests that ssASm⁶A may directly or indirectly regulate the occurrence and development of lung cancer by modifying disease-related genes and interacting proteins. We conducted functional pathway analysis on the overlapping genes, revealing significant enrichment in pathways related to lung cancer, such as positive regulation of cell migration [56], epithelial cell development [56], and protein catabolic process [57] (Additional file 1: Fig. S9). This suggests that ssASm⁶A may regulate the occurrence and development of lung cancer by affecting the function of lung-cancer-related gene pathways.

Discussion

Recent research suggests the widespread presence of ASm⁶A modifications and their impact on disease susceptibility. In this study, a novel method called M6Allele was developed for detecting ASm⁶A events using MeRIP-seq data, both in single-sample analysis and in a paired-sample comparison (differential ASm⁶A). M6Allele integrates available information to determine ASm⁶A extent in a given peak by meta-analysis across SNPs within IP and Input samples. Combining M6Allele with MeRIP-seq analysis tools

in our pipeline enables precise visualization of the transcriptome-wide ASm⁶A landscape.

Due to the absence of known phase information in most of the MeRIP-seq data, M6Allele utilizes a pseudo-phasing strategy to delineate the distribution of modified reads across various haplotypes. The pseudo-phasing strategy for inferring gene haplotypes may lead to nonconservative nominal P-values when calculating the statistical significance of ASm⁶A. To assess this issue, the GPD was introduced to adjust the statistical significance level. The performance with simulated data demonstrated the robustness of this strategy, allowing M6Allele to accurately identify significant allele-specific imbalance events.

Unlike other existing algorithms, M6Allele does not identify SNP or mutation sites associated with ASm⁶A. Instead, it employs a meta-analysis strategy at the peak level, integrating all SNP information within each peak for ASm⁶A estimation through a hierarchical Bayesian model. Using the MCMC process, the probability distribution of the odds ratio for the major allele haplotypes within each peak is sampled, constructing empirical statistical tests to identify significant ASm⁶A events. This computational approach performs well across different parameters in MeRIP-seq experiments and compares favorably with other state-of-the-art tools. Additionally, the framework of M6Allele supports both within-sample and paired-sample ASm⁶A analyses. The latter functionality allows the user to, for example, identify differential ASm⁶A in tumor versus normal comparisons, or to compare ASm⁶A changes before and after treatment. These features make M6Allele more suitable for identifying ASm⁶A events under real experimental conditions. This study applied the M6Allele to identify ASm⁶A events in pulmonary fibrosis and lung adenocarcinoma. The results demonstrated a significant association between the identified ASm⁶A genes and these conditions, revealing the potential key role of ASm⁶As in the development of these diseases. This also indicates that M6Allele can provide a reliable ASm⁶A landscape for downstream experimental research.

Although M6Allele was originally designed for MeRIP-seq experiments, it is also applicable for peak detection and differential analysis of other RIP-seq data, such as m⁷G or Ac⁴C. However, since the peak-calling tools within the M6Allele pipeline are primarily optimized for MeRIP-seq data, users can alternatively upload peak-calling results from other tools to facilitate the analysis of ASM events across various RIP-seq datasets. In this study, we employed a pseudo-phasing strategy, which may introduce some deviation in MAF values, albeit insignificantly affecting events with marked allelic imbalances. Therefore, integrating gold-standard haplotype data such as whole-genome sequencing data will be considered to enhance M6Allele's performance. Additionally, the gene dataset used to calibrate GPD for P-value correction comprises solely human genes. Nevertheless, given the homologous nature of gene expression, the P-value correction model remains applicable to studies involving other vertebrates. To ensure more precise assessments, our future endeavors will encompass a broader array of species within the M6Allele model, encompassing mice, fruit flies, yeast, and zebrafish.

Conclusions

This study showed that M6Allele is a powerful tool for detecting ASm⁶A events using MeRIP-seq data, offering significant advantages in visualizing the transcriptome-wide ASm⁶A landscape. The method's ability to handle both single-sample and paired-sample analyses provides versatility in identifying significant ASm⁶A events under various experimental conditions. Applying

M6Allele to pulmonary fibrosis and lung adenocarcinoma data highlighted its potential in uncovering the role of ASm⁶As in disease development. While the pseudo-phasing strategy and haplotype reconstruction method have some limitations, introducing GPD for P-value adjustment ensures more accurate statistical significance assessments. This study paves the way for more comprehensive studies on the interplay between m⁶A modifications and disease genetics, contributing valuable insights to the field. It sets the stage for more in-depth studies on how m⁶A modifications interact with disease genetics, providing valuable insights.

Methods

Overview of M6Allele

The comprehensive mathematical description and justification for M6Allele is provided in the Supplementary Methods. Here, we offer a summary of M6Allele and its application in this manuscript.

The core algorithm of M6Allele is comprised of three functional modules: (1) a module that infers genes with significant ASE events in RNA-seq samples (Fig. 1A); (2) a module designed to identify ASm⁶A peaks from a single MeRIP-seq sample; and (3) a module for detecting the differential ASm⁶A peak between paired samples.

Construction of the ASE determination module

We use genes as the units of ASE, defined as the combination of all exons that form individual transcript isoforms. M6Allele models the logarithm of the odds ratio of the major haplotype in a gene using a normal distribution.

However, the framework depends on specifying gene haplotypes, which may be unknown for MeRIP-seq datasets. Here, we refer to the voting-based pseudo-phasing strategy in MBASED [32] for haplotyping. When a gene contains at least one heterozygous exon SNP, we assume it to have two haplotypes. We then count the reads mapping to individual SNPs in the Input sample and define the top two highest read counts of bases as the “major” and “minor” haplotypes for that site.

For a given gene, the following notation will be used upon describing the raw input:

- n_j , total reads of the j th SNP site in the gene;
- $x_{ma,j}$, the count of reads mapping to the major haplotype in SNP _{j} ;
- $x_{0,j}$, the theoretical read counts of the major haplotype at SNP _{j} without ASE, with a default value of $0.5 \times n_j$.

Accordingly, the standardized odds ratio ρ_j of the major haplotype at an individual SNP _{j} can be represented as:

$$\rho_j = \frac{x_{ma,j}}{n_j - x_{ma,j}} / \frac{x_{0,j}}{n_j - x_{0,j}}. \quad (1)$$

The logarithm form of ρ_j is then computed as:

$$y_j = \ln(\rho_j) = \ln\left(\frac{x_{ma,j}}{n_j - x_{ma,j}}\right) - \ln\left(\frac{x_{0,j}}{n_j - x_{0,j}}\right). \quad (2)$$

Sequencing biases and subsequent analytic process such as read alignments can usually cause fluctuations in observed read counts, making them deviate from theoretical values. Therefore, it is necessary to consider these fluctuations when estimating the logarithm of odds ratios for SNPs. To address this, we have introduced the one-way normal random-effects model (REM) [33] and

assumed that each observed y_j in a gene is generated from the following process:

$$y_j \sim N(\theta_j, \sigma_j^2), \quad (3)$$

$$\theta_j \sim N(\mu, \tau^2), \quad (4)$$

$$\mu \sim \text{Uniform}(-\infty, +\infty), \quad (5)$$

$$\tau \sim \text{scale} - \text{Inv} - \chi^2(\varphi, s^2). \quad (6)$$

Notably, μ is the expected value of θ_j . Estimating μ provides the global log odds ratio for the major haplotype of the gene and serves as a measure of ASE extent. It is worth noting that we found that the two parameters of the previous distribution for τ have a negligible impact on the identification performance of allele-specific events (Additional file 1: Fig. S7). Therefore, in the subsequent analysis, we set $\varphi = 5$ and $s^2 = 10$.

Using the improved Metropolis–Hastings (M–H) sampling method based on the Markov chain Monte Carlo (MCMC) algorithm, we sample from the joint posterior distribution (the full derivation is shown in the Supplementary Methods):

$$p(\theta_1, \dots, \theta_j, \mu, \tau | y_1, y_2, \dots, y_j), \quad (7)$$

and simultaneously their marginals:

$$p(\mu, \tau | y_1, y_2, \dots, y_j) \quad (8)$$

and

$$p(\tau | y_1, y_2, \dots, y_j). \quad (9)$$

We then compute the posterior means for μ , which we denote as μ_{ASE} , as the indicator of ASE.

Construction of the ASm⁶A determination module

The ASm⁶A determination module is similar to the ASE determination module, using m⁶A peaks as the meta-analysis unit and the SNP sites covered by each modification peak for hierarchical Bayesian model construction.

For a given peak, the following notation will be used to describe this step:

$n_j^{(m)}$, the total number of reads observed at that site in the IP sample;

$x_{ma,j}^{(m)}$, represents the read count for the major haplotype of the j th SNP locus within the peak in the IP sample.

To eliminate the influence of ASE on ASm⁶A identification, we will use the previously calculated gene ASE odds ratio μ_{ASE} as the background for calculating the ASm⁶A odds ratio $\rho_j^{(m)}$, with the following equation:

$$\rho_j^{(m)} = \frac{x_{ma,j}^{(m)}}{n_j^{(m)} - x_{ma,j}^{(m)}} / e^{\mu_{\text{ASE}}}. \quad (10)$$

Furthermore, the equation for calculating the log odds ratio of ASm⁶A is as follows:

$$y_j^{(m)} = \ln(\rho_j^{(m)}) = \ln\left(\frac{x_{ma,j}^{(m)}}{n_j^{(m)} - x_{ma,j}^{(m)}}\right) - \mu_{\text{ASE}}. \quad (11)$$

Similar to the ASE module, we constructed a hierarchical Bayesian model for each peak with the following process:

$$y_j^{(m)} \sim N(\theta_j^{(m)}, \sigma_j^{(m)^2}), \quad (12)$$

$$\theta_j^{(m)} \sim N(\mu^{(m)}, \tau^{(m)^2}), \quad (13)$$

$$\mu^{(m)} \sim \text{Uniform}(-\infty, +\infty), \quad (14)$$

$$\tau^{(m)} \sim \text{scale} - \text{Inv} - \chi^2(\varphi, s^2). \quad (15)$$

Using the M–H sampling algorithm to estimate the parameters in the model, we can convert the calculated the posterior means for $\mu^{(m)}$ into the MAF of the peak to assess the tendency of allelic modification imbalance.

Construction of the paired-sample analysis module

In practical research, when samples from different groups originate from the same individual, they are referred to as paired samples—e.g., tumor and normal samples from the same patient. Researchers focus on intergroup differences not influenced by individual genetic information, such as sample-specific ASm⁶A events. However, using a pseudo-phasing strategy may cause inconsistent haplotyping between samples when identifying ASm⁶A separately for each sample, making it challenging to detect significant ASm⁶A differences accurately. To address this issue, we have introduced a paired-sample analysis feature that builds upon the single-sample ASm⁶A analysis. We describe the procedure here in terms of comparing a “tumor” sample to a “normal” sample, but the analysis can be done for any paired-samples. Initially, we identify m⁶A peaks that overlap more than 50% in length between different samples as originating from the same modification event. The differential ASm⁶A events between samples can be classified into the following scenarios:

- 1) A modification event is present in the tumor sample with allele-specificity but does not appear in the normal sample; this is classified as a gain ASm⁶A event in the tumor sample.
- 2) Conversely, it is considered a loss ASm⁶A event in the tumor sample.
- 3) Another modification event is identified with allele specificity in both tumor and normal samples, but shows differing major haplotypes; this is labeled as a gain event in tumor samples.
- 4) A modification event that shows allele specificity in both tumor and normal samples, with the same major m⁶A haplotype will be assessed for the significance of intersample differences using a hierarchical Bayesian model to estimate the odds ratio of the major m⁶A haplotype. We consider the consensus heterozygous SNP sites within the combined regions of these peaks as available sites for the downstream analysis, ensuring consistent haplotyping between the two samples. For each SNP site, the odds ratio calculation formula is constructed as shown below:

$$\rho_j^s = \frac{\rho_{\text{tumor},j}}{\rho_{\text{normal},j}} \quad (16)$$

where $\rho_{\text{tumor},j} = \frac{y_{\text{tumor},j}}{n_{\text{tumor},j} - y_{\text{tumor},j}} / e^{\mu_{\text{tumor},b}}$ and $\rho_{\text{normal},j} = \frac{y_{\text{normal},j}}{n_{\text{normal},j} - y_{\text{normal},j}} / e^{\mu_{\text{normal},b}}$.

Under the null hypothesis of no sample-specific ASm⁶A event occurring, we consider $\rho_{\text{tumor},j} = \rho_{\text{normal},j}$. We then construct a Bayesian model similarly with the single-sample analysis for the M–H sampling giving the expected value of the natural logarithm of ρ_j^s .

Significance threshold for ASm⁶A/ASE events

The hierarchical Bayesian models merely compute tendencies of allele-specific events. To identify significant allele-specific events,

we need to construct a testing model. Here, we developed a threshold calculation algorithm based on extreme value theory to assess the significance of allele-specific events. Details of the threshold calculation algorithm can be found in the Supplementary Methods.

To distinguish significant allele-specific events, we are required to obtain the minor allele frequency (MAF) distribution under the null hypothesis condition. Due to the lack of eligible real MeRIP-seq data meeting the criteria, we need to simulate sequencing data to obtain the read counts for major and minor alleles of SNPs without significant ASE or ASm⁶A events. Since previous studies commonly fit the read distribution with a negative binomial distribution (NBD), we also introduce it here to fit the read count distribution for individual SNPs captured by sequencing.

When the total read count for SNP_i is N_i , the read count x_{ij} for individual allele j (j can be 0 or 1) covering each SNP locus is assumed as follows:

$$x_{ij} \sim NB(\omega_i, k). \quad (17)$$

Here, ω_i represents the theoretical read count of one haplotype at a SNP site without allele-specific events, so it can be calculated using $0.5N_i$. In addition, k is the dispersion parameter.

Next, we need to estimate k using appropriate sequencing data. Since most heterozygous somatic mutations on diploid genomes typically involve only one chromosome, genome-wide sequencing data for detecting genomic mutations theoretically lack allele imbalance and are suitable as background data for estimating k . To evaluate individual heterogeneity in actual sequencing data and determine the dispersion of read counts, we obtained whole-genome sequencing (WGS) data from the 1000 Genomes database. We then tallied the read counts N_i at SNP_i along with the read counts x_{i0} and x_{i1} for the alleles. Since each SNP from different individuals can be considered independently distributed, we integrated all the N_i and x_{ij} using maximum-likelihood estimation to estimate the dispersion parameter k .

Subsequently, we simulated the total read counts for each SNP on every gene/peak as the parameter ω_i of the NBD. Given the varied gene expression patterns in the transcriptome, we established gene-specific FPKM distributions to enhance the fidelity of our simulated data reflecting true gene expression. We collected FPKM values for all genes from The Cancer Genome Atlas Program (TCGA) [58] and fitted their distributions for each gene using the Python package *fitter* (<https://pypi.org/project/fitter/>). Genes were classified into six categories with the distribution type of FPKM according to previous research [59]. To facilitate computation, we refitted the overall distribution of FPKM for each category and sampled from these distributions to simulate FPKM values for each gene within its respective class. Simultaneously, by simulating the library size of sequencing data, we further calculated the total read count N'_i for each SNP on the gene based on the simulated FPKM and gene length.

Based on the dispersion parameter k and N'_i , we derived the NBD for the allelic reads of each SNP within the gene/peak. By sampling from the NBD, we simulated the counts of reads for major and minor alleles of every SNP, and obtained MAF for each gene/peak using M6Allele.

Given the rarity of allele-specific events, we assume they follow a tail distribution in genomic data. Thus, we introduced the generalized Pareto distribution (GPD), which accurately models the tails of various distributions. In the categorization of different gene expression patterns, we estimated the tail distribution of MAF under the null hypothesis and computed the statistical significance thresholds.

The implementation and integration of M6Allele

We implemented the single- and paired-sample analyses described above in a JAR package called M6Allele. To enhance users' convenience, we provided a comprehensive pipeline for ASm⁶A analysis using Docker. This pipeline integrates tools such as FastQC (RRID:SCR_014583), fastp (RRID:SCR_016962) [60], STAR (RRID:SCR_004463) [61], VARSCAN (RRID:SCR_006849) [62], GATK (RRID:SCR_001876) [63], and MetPeak (RRID:SCR_026533) [64] for quality control, alignment, SNP calling, and m⁶A peak calling. While MetPeak was used as the default peak-calling tool in this study, we have tested other peak-calling tools, such as TRESS [65] and exomePeak2 [66], and MACS3 [67], and confirmed that they are also compatible with M6Allele. By providing FASTQ sequencing files, gene annotation GTF files, and reference genome fasta files, users can automatically calculate allele-specific events for both gene expressions and m⁶A modifications. The pipeline generates reports on MAF and ASE/ASm⁶A P-values for each allele-specific event.

MeRIP-seq data collection and alignment

MeRIP-seq raw sequencing reads for pulmonary fibrosis and lung carcinoma were downloaded from the NCBI Gene Expression Omnibus [68] (GEO (<https://www.ncbi.nlm.nih.gov/geo/>); accession numbers GSE164151, GSE198288). FastX_Trimmer (version 0.0.13) and FastQC (version 0.11.9) was used to trim adaptors and control read quality, respectively. Then, the clean reads were mapped to the human genome (GRCh38) using STAR [61] (version 2.7.6.a) with parameters set as `–twopassMode Basic`. SAMtools [69] was then utilized to filter for uniquely aligned sequences or select the highest-scoring alignment from multiple alignments.

Variant calling from the input sample of MeRIP-seq data

VarScan [62] (version 2.3.9) was used to detect SNPs with a minimum VAF value of 0.05. Following this, BCFtools [69] (version 1.2.1) was employed to flag SNP positions with a reference allele depth <2 bp or within 3 bp of an indel. Then, VCFtools [70] (version 0.1.17) was applied to filter out the flagged positions. The variants were retained if they matched the criteria: neither were found in UCSC RepeatMasker microsatellites [71] nor in RNA editing sites (RADAR database [34]) but were contained in the dbSNP database [38] or the 1000 Genomes. Then we count the reads on the two alleles for each SNP. Only those variants that satisfied the minimum mapping reads on both alleles were considered as reliable candidate heterozygous sites (each allele ≥ 2 , the sum of two alleles ≥ 10 [29]).

m⁶A peak calling from MeRIP-seq data

To obtain m⁶A modification peaks, we utilized MetPeak [64] (version 1.1) for peak calling with default parameter settings. By comparing with variant information, only m⁶A peaks that contain variants were retained for allele-specific methylation analysis.

Comparison with the other ASE or ASm⁶A identification methods

In comparing ASE identification methods, we utilized GeneiASE [43] and MBASED [32] to identify genes exhibiting significant ASE in simulated RNA-seq data. Leveraging the settings of true ASE events in the simulated data, we computed metrics such as Precision, Recall, FDR, F0.5, and F1 for the results obtained

from GeneiASE, MBASED, and M6Allele, facilitating a thorough comparison.

For the comparison of ASm⁶A identification tools, we executed the methods of ASPRIN and Cao et al. according to their tutorials, adhering to the default parameter estimates as suggested by the authors. As these tools analyze individual SNP sites, we aligned SNPs associated with ASm⁶A identified by these tools with m⁶A peaks. SNPs not aligning with the regions of m⁶A peaks were excluded from the analysis. If any SNP within a peak was identified as having ASm⁶A modification by ASPRIN or Cao et al.'s algorithm, that peak was classified as ASm⁶A modified, resulting in a positive outcome; otherwise, it was considered negative. Based on this strategy, we can get the accuracy of the prediction results for each peak and calculate the true-positive rate (TPR) and the false-positive rate (FPR).

Gene ontology enrichment analysis

We performed gene ontology enrichment analysis on genes with ASE or ASm⁶A modifications using Metascape [44] with the "GO Biological Process" pathway dataset. A significance level of $P < 0.05$ was chosen as the threshold for statistical significance. Following this, we imported the GO pathway enrichment results into Cytoscape [72] and utilized the ClueGO [73] plugin to visualize the pathway networks.

Cell lines and cell culture

The human monocytic THP-1 cell line (#TIB-202), originally purchased from American Type Culture Collection (ATCC) by Dr Shouheng Jin and provided for this study, was cultured in RPMI 1640 medium (Gibco, catalog number C22400500BT) supplemented with 10% FBS and 1% glutamine.

MeRIP sequencing

Total RNA was extracted using TRIzol reagent (Invitrogen, USA) and assessed for quality using a NanoDrop and Bioanalyzer. Poly(A)-tailed RNA was purified using Dynabeads Oligo(dT)25 (Thermo Fisher, USA) and fragmented at 86°C for 7 min. Fragmented RNA was incubated with an m⁶A-specific antibody (catalog number 202,003, Synaptic Systems, Germany) in IP buffer to enrich m⁶A-modified RNA. The RNA was reverse-transcribed into cDNA and converted into double-stranded DNA, followed by adapter ligation and size selection using AMPure XP beads. Libraries were amplified by PCR and sequenced on an Illumina NovaSeq 6000 platform (LC-Bio Technology Co., Ltd., Hangzhou, China) in paired-end 150 bp mode.

Sanger sequencing

To validate potential ASm⁶A events, we initially selected 15 positive and 15 negative candidate sites based on M6Allele's predictions from the THP-1 cell line MeRIP-seq data. To ensure that the expression levels of the transcripts at the selected sites are sufficient for the validation, we filtered the sites by calculating the total read counts in the Input samples and excluded sites with fewer than 25 reads (Additional file 2: Table S4). As a result, a total of 6 positive and 14 negative sites were used for validation. Primers targeting these sites were designed for both Input and IP samples, with detailed primer sequences provided in Additional file 2: Table S3. PCR products were gel-purified and subjected to Sanger sequencing. The sequencing chromatograms were processed using EditR software [74] to calculate the proportions of different nucleotides at the selected sites. Odds ratios for the major allele were calculated to compare nu-

cleotide proportions between IP and input samples, with sites having an odds ratio greater than 1.2 classified as positive ASm⁶A events.

Availability of Source Code and Requirements

Project Name: M6Allele

Project Homepage: <https://github.com/RenLabBioinformatics/M6Allele>

Operating System(s): Platform independent

Programming Language: Java

Other Requirements: This pipeline integrates multiple tools, including FastQC, fastp [60], STAR [61], VarScan [62], GATK [63], and MetPeak [64], for quality control, alignment, SNP calling, and m⁶A peak identification. All dependencies are prepackaged in the provided Docker image [32], and the workflow is registered on WorkflowHub [75].

License: MIT License

RRID:SCR_026077

Bio.tools ID: biotools:m6allele

The Docker image file of M6Allele, containing the JAR file and all necessary dependencies, can be downloaded from [32]. Comprehensive installation and usage instructions are available on [76].

Additional Files

Supplementary Methods and Supplementary Figure S1-S10.

Fig. S1. Simulation data generation workflow.

Fig. S2. Comparison of the performance of M6Allele, ASPRIN, and Cao et al. in different metrics.

Fig. S3. Performance comparison of different ASm⁶A detection tools across various evaluation metrics on simulated datasets.

Fig. S4. Comparison of computational time for ASE and ASm⁶A detection across different tools and sample sizes.

Fig. S5. IGV visualizations on simulated data.

Fig. S6. IGV visualizations of differential ASm⁶A.

Fig. S7. ASE gene analysis results on pulmonary fibrosis dataset.

Fig. S8. Enrichment and pathway analysis of ASm⁶A and ASE genes in pulmonary fibrosis.

Fig. S9. Pathway analysis of overlapping ssASm⁶A-modified genes and their interactors enriched in lung cancer-related pathways.

Fig. S10. Impact of scaled inverse chi-squared prior parameters on predicted major allele frequency.

Table S1. Results of model identification of ASE events on simulated data.

Table S2. Identification of ASM events in THP-1 cell line MeRIP-seq data using M6Allele, ASPRIN, and Cao et al.

Table S3. Primers information of ASm⁶A sites used for Sanger sequencing experiments.

Table S4. Base composition of ASm⁶A sites in input and IP samples analyzed by Sanger sequencing.

Table S5. Identification of ASE events in pulmonary fibrosis dataset through single-sample analysis using M6Allele.

Table S6. Identification of ASM events in pulmonary fibrosis dataset through single-sample analysis using M6Allele.

Table S7. Identification of ASM events in lung adenocarcinoma dataset through paired-sample analysis using M6Allele.

Table S8. ssASm⁶A genes in the lung adenocarcinoma dataset.

Abbreviations

ASE: Allele-specific gene expression; ASm6A: allele-specific m⁶A; ASPRIN: Allele-Specific Protein_RNA Interaction; ATCC: American Type Culture Collection; AUC: area under the ROC curve; BH: Benjamini_Hochberg; circRNAs: circular RNAs; FPR: false-positive rate; GEO: Gene Expression Omnibus; GPD: generalized Pareto distribution; m6A: RNA N⁶-methyladenosine; MAF: the frequency of the major allele; M-H: Metropolis_Hastings; miRNAs: microRNAs; mRNAs: messenger RNAs; NBD: negative binomial distribution; NFkB: nuclear factor kappa B; REM: random-effects model; ssASm6A: sample-specific ASm6A; TCGA: The Cancer Genome Atlas Program; TPR: true-positive rate.

Acknowledgments

We thank Dr. Shouheng Jin for providing the THP-1 cell line used in this study.

Author Contributions

X.L. and Y.X. conceived the project. Y.Z., L.T., and S.Z. developed the methodology and implemented the method. B.H. and Z.Z. helped with the design of methodology. X.L., Y.X., and J.R. wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported by National Key Research and Development Program of China (2023YFC2705900); the National Natural Science Foundation of China (32200542, 82301233); the Young Elite Scientists Sponsorship Program by Guangzhou Association for Science and Technology (QT-2023-045); the Guangdong Province Excellent Youth Team Project (2024B1515040009); and the Discipline Training, Innovation and Quality Improvement Engineering Team Project of Guangdong Pharmaceutical University (2024QZ02).

Data Availability

The raw THP-1 cell line MeRIP-seq data used in this study have been deposited in the Gene Expression Omnibus (GEO) under accession code GSE289760, with the corresponding BioProject accession PRJNA1224735, and in the Genome Sequence Archive [77] in National Genomics Data Center [78], China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences under NGDC_BioProject: PRJCA034320. MeRIP-seq raw sequencing data for pulmonary fibrosis and lung carcinoma were obtained from GEO with accession numbers GSE164151 and GSE198288. The raw Sanger sequencing results generated in this study have been deposited in GigaDB and are available in [79]. In addition, GigaDB hosts an archival copy of the analysis code, the M6Allele software package, example datasets for ASE and ASM detection, simulation and real experimental data, as well as supporting tables summarizing key findings.

Competing Interests

The authors have declared that no competing interests exist.

Ethics Approval and Consent to Participate

This study includes the use of publicly available datasets obtained from open-access databases as well as newly generated sequencing data from the THP-1 cell line. The human monocytic THP-1 cell line (#TIB-202) was originally purchased from the American Type Culture Collection (ATCC) by Dr Shouheng Jin, who kindly provided it for use in this study. The cells were cultured following standard protocols. As the THP-1 cell line is not classified as human research material, no additional ethics approval or written consent was required for this study.

References

1. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet.* 2010;11:533–38. <https://doi.org/10.1038/nrg2815>.
2. Xu Q, Xiang Y, Wang Q, et al. SETD2 regulates the maternal epigenome, genomic imprinting and embryonic development. *Nat Genet.* 2019;51:844–56. <https://doi.org/10.1038/s41588-019-0398-7>.
3. Bonthuis PJ, Huang WC, Stacher Horndli CN, et al. Noncanonical genomic imprinting effects in offspring. *Cell Rep.* 2015;12:979–91. <https://doi.org/10.1016/j.celrep.2015.07.017>.
4. Sveen A, Johannessen B, Eilertsen IA, et al. The expressed mutational landscape of microsatellite stable colorectal cancers. *Genome Med.* 2021;13:142. <https://doi.org/10.1186/s13073-021-00955-2>.
5. Gendrel AV, Marion-Poll L, Katoh K, et al. Random monoallelic expression of genes on autosomes: parallels with X-chromosome inactivation. *Semin Cell Dev Biol.* 2016;56:100–10. <https://doi.org/10.1016/j.semcdb.2016.04.007>.
6. Reinius B, Sandberg R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat Rev Genet.* 2015;16:653–64. <https://doi.org/10.1038/nrg3888>.
7. van Ekelburg YS, Hornslien KS, Van Hautegeem T, et al. Spatial and temporal regulation of parent-of-origin allelic expression in the endosperm. *Plant Physiol.* 2023;191:986–1001. <https://doi.org/10.1093/plphys/kiac520>.
8. Barlow DP, Bartolomei MS. Genomic imprinting in mammals. *Cold Spring Harb Perspect Biol.* 2014;6:a018382. <https://doi.org/10.1101/cshperspect.a018382>.
9. Kravitz SN, Gregg C. New subtypes of allele-specific epigenetic effects: implications for brain development, function and disease. *Curr Opin Neurobiol.* 2019;59:69–78. <https://doi.org/10.1016/j.conb.2019.04.012>.
10. Sigurdsson MI, Saddic L, Heydarpour M, et al. Allele-specific expression in the human heart and its application to postoperative atrial fibrillation and myocardial ischemia. *Genome Med.* 2016;8:127. <https://doi.org/10.1186/s13073-016-0381-1>.
11. Gyorgy B, Nist-Lund C, Pan B, et al. Allele-specific gene editing prevents deafness in a model of dominant progressive hearing loss. *Nat Med.* 2019;25:1123–30. <https://doi.org/10.1038/s41591-019-0500-9>.
12. Sen A, Huo Y, Elster J, et al. Allele-specific expression reveals genes with recurrent cis-regulatory alterations in high-risk neuroblastoma. *Genome Biol.* 2022;23:71. <https://doi.org/10.1186/s13059-022-02640-y>.
13. Shetty A, Seo JH, Bell CA, et al. Allele-specific epigenetic activity in prostate cancer and normal prostate tissue implicates prostate cancer risk mechanisms. *Am Hum*

- Genet. 2021;108:2071–85. <https://doi.org/10.1016/j.ajhg.2021.09.008>.
14. Guo Y, Feng YF, Yang GG, et al. Allele-specific DNA methylation and gene expression during shoot organogenesis in tissue culture of hybrid poplar. *Hortic Res.* 2024;11:uhae027. <https://doi.org/10.1093/hr/uhae027>.
15. Xuan A, Song Y, Bu C, et al. Changes in DNA methylation in response to 6-benzylaminopurine affect allele-specific gene expression in *Populus tomentosa*. *Int J Mol Sci.* 2020;21:2117. <https://doi.org/10.3390/ijms21062117>.
16. Zhang Y, Rohde C, Reinhardt R, et al. Non-imprinted allele-specific DNA methylation on human autosomes. *Genome Biol.* 2009;10:R138. <https://doi.org/10.1186/gb-2009-10-12-r138>.
17. Zheng HX, Zhang XS, Sui N. Advances in the profiling of N⁶-methyladenosine (m⁶A) modifications. *Biotechnol Adv.* 2020;45:107656. <https://doi.org/10.1016/j.biotechadv.2020.107656>.
18. Han X, Guo J, Fan Z. Interactions between m6A modification and miRNAs in malignant tumors. *Cell Death Dis.* 2021;12:598. <https://doi.org/10.1038/s41419-021-03868-5>.
19. Feng ZH, Liang YP, Cen JJ, et al. m6A-immune-related lncRNA prognostic signature for predicting immune landscape and prognosis of bladder cancer. *J Transl Med.* 2022;20:492. <https://doi.org/10.1186/s12967-022-03711-1>.
20. Du A, Li S, Zhou Y, et al. M6A-mediated upregulation of circMDK promotes tumorigenesis and acts as a nanotherapeutic target in hepatocellular carcinoma. *Mol Cancer.* 2022;21:109. <https://doi.org/10.1186/s12943-022-01575-z>.
21. Liu H, Zheng J, Liao A. The regulation and potential roles of m6A modifications in early embryonic development and immune tolerance at the maternal–fetal interface. *Front Immunol.* 2022;13:988130. <https://doi.org/10.3389/fimmu.2022.988130>.
22. Yang Z, Cai Z, Yang C, et al. ALKBH5 regulates STAT3 activity to affect the proliferation and tumorigenicity of osteosarcoma via an m6A-YTHDF2-dependent manner. *EBioMedicine.* 2022;80:104019. <https://doi.org/10.1016/j.ebiom.2022.104019>.
23. Kasowitz SD, Ma J, Anderson SJ, et al. Nuclear m6A reader YTHDC1 regulates alternative polyadenylation and splicing during mouse oocyte development. *PLoS Genet.* 2018;14:e1007412. <https://doi.org/10.1371/journal.pgen.1007412>.
24. Yin H, Zhang X, Yang P, et al. RNA m6A methylation orchestrates cancer growth and metastasis via macrophage reprogramming. *Nat Commun.* 2021;12:1394. <https://doi.org/10.1038/s41467-021-21514-8>.
25. Azzam SK, Alsafar H, Sajini AA. FTO m6A demethylase in obesity and cancer: implications and underlying molecular mechanisms. *Int J Mol Sci.* 2022;23:3800. <https://doi.org/10.3390/ijms23073800>.
26. Xiong X, Hou L, Park YP, et al. Genetic drivers of m⁶A methylation in human brain, lung, heart and muscle. *Nat Genet.* 2021;53:1156–65. <https://doi.org/10.1038/s41588-021-00890-3>.
27. Olazagoitia-Garmendia A, Rojas-Marquez H, Sebastian-delacruz M, et al. m⁶A methylated long noncoding RNA LOC339803 regulates intestinal inflammatory response. *Adv Sci.* 2024;11:e2307928. <https://doi.org/10.1002/adv.202307928>.
28. Olazagoitia-Garmendia A, Zhang L, Mera P, et al. Gluten-induced RNA methylation changes regulate intestinal inflammation via allele-specific XPO1 translation in epithelial cells. *Gut.* 2022;71:68–76. <https://doi.org/10.1136/gutjnl-2020-322566>.
29. Cao S, Zhu H, Cui J, et al. Allele-specific RNA N⁶-methyladenosine modifications reveal functional genetic variants in human tissues. *Genome Res.* 2023;33:1369–80. <https://doi.org/10.1101/gr.277704.123>.
30. Bahrami-Samani E, Xing Y. Discovery of allele-specific protein–RNA interactions in human transcriptomes. *Am Hum Genet.* 2019;104:492–502. <https://doi.org/10.1016/j.ajhg.2019.01.018>.
31. Guk JY, Jang MJ, Choi JW, et al. De novo phasing resolves haplotype sequences in complex plant genomes. *Plant Biotechnol J.* 2022;20:1031–41. <https://doi.org/10.1111/pbi.13815>.
32. m6allelepipe. <https://renlab.oss-cn-shenzhen.aliyuncs.com/M6Allele/m6allelepipe.tar.gz>. Accessed 7 March 2024.
33. Castel SE, Levy-Moonshine A, Mohammadi P, et al. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* 2015;16:195. <https://doi.org/10.1186/s13059-015-0762-6>.
34. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucl Acids Res.* 2014;42:D109–13. <https://doi.org/10.1093/nar/gkt996>.
35. Mayba O, Gilbert HN, Liu J, et al. MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.* 2014;15:405. <https://doi.org/10.1186/s13059-014-0405-3>.
36. Borenstein M, Hedges LV, Higgins JP, et al. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Method.* 2010;1:97–111. <https://doi.org/10.1002/jrsm.12>.
37. 1000 Genomes. <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data>. Accessed 20 December 2022.
38. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29:308–11. <https://doi.org/10.1093/nar/29.1.308>.
39. Wang C, Chen G. A new hybrid estimation method for the generalized pareto distribution. *Commun Stat Theory Methods.* 2016;45:4285–94. <https://doi.org/10.1080/03610926.2014.919399>.
40. Albaradei S, Thafar M, Alsaedi A, et al. Machine learning and deep learning methods that use omics data for metastasis prediction. *Comput Struct Biotechnol J.* 2021;19:5008–18. <https://doi.org/10.1016/j.csbj.2021.09.001>.
41. Frazee AC, Jaffe AE, Langmead B, et al. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics.* 2015;31:2778–84. <https://doi.org/10.1093/bioinformatics/btv272>.
42. Edsgard D, Iglesias MJ, Reilly SJ, et al. GeneiASE: detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Sci Rep.* 2016;6:21134. <https://doi.org/10.1038/srep21134>.
43. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inform Process Manage.* 2009;45:427–37. <https://doi.org/10.1016/j.ipm.2009.03.002>.
44. Zhou Y, Zhou B, Pache L, et al. Metascope provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019;10:1523. <https://doi.org/10.1038/s41467-019-09234-6>.
45. Pinero J, Ramirez-Anguita JM, Sauch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 2020;48:D845–55. <https://doi.org/10.1093/nar/gkz1021>.
46. Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 2023;51:D638–46. <https://doi.org/10.1093/nar/gkac1000>.
47. Wang Z, Liu Y, Chen F, et al. Feasibility and mechanism analysis of Reduning in the prevention of sepsis-induced pulmonary fibrosis. *Front Pharmacol.* 2022;13:1079511. <https://doi.org/10.3389/fphar.2022.1079511>.

48. Rajesh R, Atallah R, Barnthaler T. Dysregulation of metabolic pathways in pulmonary fibrosis. *Pharmacol Ther.* 2023;246:108436. <https://doi.org/10.1016/j.pharmthera.2023.108436>.
49. Guan S, Zhou J. CXCR7 attenuates the TGF-beta-induced endothelial-to-mesenchymal transition and pulmonary fibrosis. *Mol Biosyst.* 2017;13:2116–24. <https://doi.org/10.1039/C7MB00247E>.
50. Grimminger F, Gunther A, Vancheri C. The role of tyrosine kinases in the pathogenesis of idiopathic pulmonary fibrosis. *Eur Respir J.* 2015;45:1426–33. <https://doi.org/10.1183/09031936.00149614>.
51. Scruggs AM, Koh HB, Tripathi P, et al. Loss of CDKN2B promotes fibrosis via increased fibroblast differentiation rather than proliferation. *Am J Respir Cell Mol Biol.* 2018;59:200–14. <https://doi.org/10.1165/rcmb.2017-0298OC>.
52. Liu Y, Yang D, Liu T, et al. N6-methyladenosine-mediated gene regulation and therapeutic implications. *Trends Mol Med.* 2023;29:454–67. <https://doi.org/10.1016/j.molmed.2023.03.005>.
53. Li K, Peng ZY, Wang R, et al. Enhancement of TKI sensitivity in lung adenocarcinoma through m6A-dependent translational repression of wnt signaling by circ-FBXW7. *Mol Cancer.* 2023;22:103. <https://doi.org/10.1186/s12943-023-01811-0>.
54. Fang H, Sun Q, Zhou J, et al. m⁶A methylation reader IGF2BP2 activates endothelial cells to promote angiogenesis and metastasis of lung adenocarcinoma. *Mol Cancer.* 2023;22:99. <https://doi.org/10.1186/s12943-023-01791-1>.
55. Zhang JX, Huang PJ, Wang DP, et al. m⁶A modification regulates lung fibroblast-to-myofibroblast transition through modulating KCNH6 mRNA translation. *Mol Ther.* 2021;29:3436–48. <https://doi.org/10.1016/j.ymthe.2021.06.008>.
56. Guo L, Liu Z, Tang X. Overexpression of SLFN5 induced the epithelial-mesenchymal transition in human lung cancer cell line A549 through beta-catenin/snail/E-cadherin pathway. *Eur J Pharmacol.* 2019;862:172630. <https://doi.org/10.1016/j.ejphar.2019.172630>.
57. Wang X, Chen X, Liu H. Expression and bioinformatics-based functional analysis of UAP1 in lung adenocarcinoma. *CMAR.* 2020;12:12111–21. <https://doi.org/10.2147/CMAR.S282238>.
58. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 2013;45:1113–20. <https://doi.org/10.1038/ng.2764>.
59. de Torrente L, Zimmerman S, Suzuki M, et al. The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. *BMC Bioinf.* 2020;21:562. <https://doi.org/10.1186/s12859-020-03892-w>.
60. Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* 2018;34:i884–90. <https://doi.org/10.1093/bioinformatics/bty560>.
61. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
62. Koboldt DC, Larson DE, Wilson RK. Using VarScan 2 for germline variant calling and somatic mutation detection. *Curr Protoc Bioinform.* 2013;44:15.4.1–17. <https://doi.org/10.1002/0471250953.bi1504s44>.
63. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–303. <https://doi.org/10.1101/gr.107524.110>.
64. Cui X, Meng J, Zhang S, et al. A novel algorithm for calling mRNA m6A peaks by modeling biological variances in MeRIP-seq data. *Bioinformatics.* 2016;32:i378–85. <https://doi.org/10.1093/bioinformatics/btw281>.
65. Guo Z, Shafik AM, Jin P, et al. Differential RNA methylation analysis for MeRIP-seq data under general experimental design. *Bioinformatics.* 2022;38:4705–12. <https://doi.org/10.1093/bioinformatics/btac601>.
66. Meng J, Lu Z, Liu H, et al. A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/bioconductor package. *Methods.* 2014;69:274–81. <https://doi.org/10.1016/j.ymeth.2014.06.008>.
67. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9:R137. <https://doi.org/10.1186/gb-2008-9-9-r137>.
68. Clough E, Barrett T. The Gene Expression Omnibus Database. *Methods Mol Biol.* 2016;1418:93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.
69. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021;10:giab008. <https://doi.org/10.1093/gigascience/giab008>.
70. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–58. <https://doi.org/10.1093/bioinformatics/btr330>.
71. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinform.* 2009;25:4.10.1–14. <https://doi.org/10.1002/0471250953.bi0410s25>.
72. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13:2498–504. <https://doi.org/10.1101/gr.1239303>.
73. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plugin to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics.* 2009;25:1091–93. <https://doi.org/10.1093/bioinformatics/btp101>.
74. Kluesner MG, Nedveck DA, Lahr WS, et al. EditR: a method to quantify base editing from sanger sequencing. *CRISPR J.* 2018;1:239–50. <https://doi.org/10.1089/crispr.2018.0014>.
75. Zhang Y, Tang L, Zhi S, et al. M6Allele. WorkflowHub. 2025. <https://doi.org/10.48546/WORKFLOWHUB.WORKFLOW.1223.1>.
76. M6Allele. <https://github.com/RenLabBioinformatics/M6Allele>. Accessed 15 June 2024.
77. Chen T, Chen X, Zhang S, et al. The Genome Sequence Archive family: toward explosive data growth and diverse data types. *Genom Proteom. Bioinform.* 2021;19:578–83. <https://doi.org/10.1016/j.gpb.2021.08.001>.
78. CNCR-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res.* 2024;52:D18–32. <https://doi.org/10.1093/nar/gkad1078>.
79. Zhang Y, Tang L, Zhi S, et al. Supporting data for “M6Allele: a toolkit for detection of allele-specific RNA N6-methyladenosine modifications”. *GigaScience Database.* 2025. <https://doi.org/10.5524/102670>.