nature genetics

Widespread adenine N6-methylation of active genes in fungi

Stephen J Mondo[1,12], Richard O Dannebaum[1,11,12], Rita C Kuo[1], Katherine B Louie[1], Adam J Bewick[2], Kurt LaButti[1], Sajeet Haridas[1], Alan Kuo[1], Asaf Salamov[1], Steven R Ahrendt[1,3], Rebecca Lau[1], Benjamin P Bowen[1], Anna Lipzen[1], William Sullivan[1], Bill B Andreopoulos[1], Alicia Clum[1], Erika Lindquist[1], Christopher Daum[1], Trent R Northen[1], Govindarajan Kunde-Ramamoorthy[1,11], Robert J Schmitz[2], Andrii Gryganskyi[4], David Culley[5], Jon Magnuson[5], Timothy Y James[6], Michelle A O'Malley[7], Jason E Stajich[8], Joseph W Spatafora[9], Axel Visel[1,10] & Igor V Grigoriev[1,3]

N6-methyldeoxyadenine (6mA) is a noncanonical DNA base modification present at low levels in plant and animal genomes[1–4], but its prevalence and association with genome function in other eukaryotic lineages remains poorly understood. Here we report that abundant 6mA is associated with transcriptionally active genes in early-diverging fungal lineages[5]. Using single-molecule long-read sequencing of 16 diverse fungal genomes, we observed that up to 2.8% of all adenines were methylated in early-diverging fungi, far exceeding levels observed in other eukaryotes and more derived fungi. 6mA occurred symmetrically at ApT dinucleotides and was concentrated in dense methylated adenine clusters surrounding the transcriptional start sites of expressed genes; its distribution was inversely correlated with that of 5-methylcytosine. Our results show a striking contrast in the genomic distributions of 6mA and 5-methylcytosine and reinforce a distinct role for 6mA as a gene-expression-associated epigenomic mark in eukaryotes.

While 5-methylcytosine (5mC) has been well-established as an epigenomic mark in eukaryotes[6], 6mA has been predominantly appreciated as a modification of prokaryotic genomes[7] and eukaryotic RNA[8]. However, genomic 6mA has recently been demonstrated to play crucial roles in both gene and transposon regulation, where it either (i) suppresses expression of transposable elements in animals during development[2–4] or (ii) is associated with promoters of actively expressed genes and is involved in nucleosome positioning in algae[1]. These studies demonstrate that while 6mA is present in eukaryotes, it may serve distinct genomic functions across diverse phylogenetic groups.

In the present study we explored adenine methylation across Fungi, a sister kingdom to Metazoa. The kingdom is estimated at ~1 billion years old[5,9] and is currently taxonomically organized into eight phyla. The six phyla at the base of the fungal kingdom belong to a relatively understudied group collectively referred to as 'early-diverging' fungi[10], while the remaining two phyla comprise Dikarya, a lineage that diverged from other fungi approximately 500 million years ago[5,9] (**Fig. 1a**). Fungi have long served as model organisms for the study of basic eukaryotic biology, including DNA base modifications such as 5mC (ref. 6). In fungi, 5mC is frequently involved in genome defense against transposable-element proliferation[6,11] and occurs symmetrically on palindromic CpGs of opposing DNA strands. While primarily restricted to repeats in fungi, in other eukaryotes, 5mC is also found within genes but is commonly excluded from CpG-rich regions surrounding gene promoters. Notably, this distribution contrasts with adenine methylation in algae (*Chlamydomonas reinhardtii*), where promoter regions appear to be enriched in methylated adenines[1].

Here we used single-molecule real-time (SMRT) sequencing[12] to decode and analyze 16 genomes from across the fungal phylogeny for presence of adenine methylation (**Supplementary Table 1** and **Supplementary Fig. 1**). Our sampling covered both phyla of Dikarya (n = 6 samples) and all phyla of early-diverging fungi except Cryptomycota and Microsporidia (n = 10 samples; **Fig. 1a**). In Dikarya, we observed 6mA levels broadly similar to those previously reported in model eukaryotic genomes[1–4], ranging from 0.048% (*Leucosporidiella creatinivora;* Pucciniomycotina) to 0.21% (*Protomyces lactucaedebilis*; Taphrinomycotina). In contrast, we found high levels of 6mA in some early-diverging fungi, where up to 2.8% (*Hesseltinella vesiculosa;* Mucoromycotina) of all adenines were methylated (**Fig. 1a,b** and **Supplementary Table 1**).

To confirm the abundance and accuracy of SMRT-detected 6mA in early-diverging fungi, we performed both mass spectrometry analysis

[1]US Department of Energy Joint Genome Institute, Walnut Creek, California, USA. [2]Department of Genetics, University of Georgia, Athens, Georgia, USA. [3]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, USA. [4]L. F. Lambert Spawn Co, Coatesville, Pennsylvania, USA. [5]Pacific Northwest National Laboratory, Richland, Washington, USA. [6]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA. [7]Department of Chemical Engineering, University of California, Santa Barbara, Santa Barbara, California, USA. [8]Department of Plant Pathology and Microbiology, University of California, Riverside, Riverside, California, USA. [9]Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA. [10]School of Natural Sciences, University of California, Merced, Merced, California, USA. [11]Present addresses: Roche Sequencing Solutions Inc, Pleasanton, California, USA (R.O.D.); The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA (G.K.R.). [12]These authors contributed equally to this work. Correspondence should be addressed to I.V.G. (ivgrigoriev@lbl.gov).

Received 26 July 2016; accepted 7 April 2017; published online 8 May 2017; doi:10.1038/ng.3859

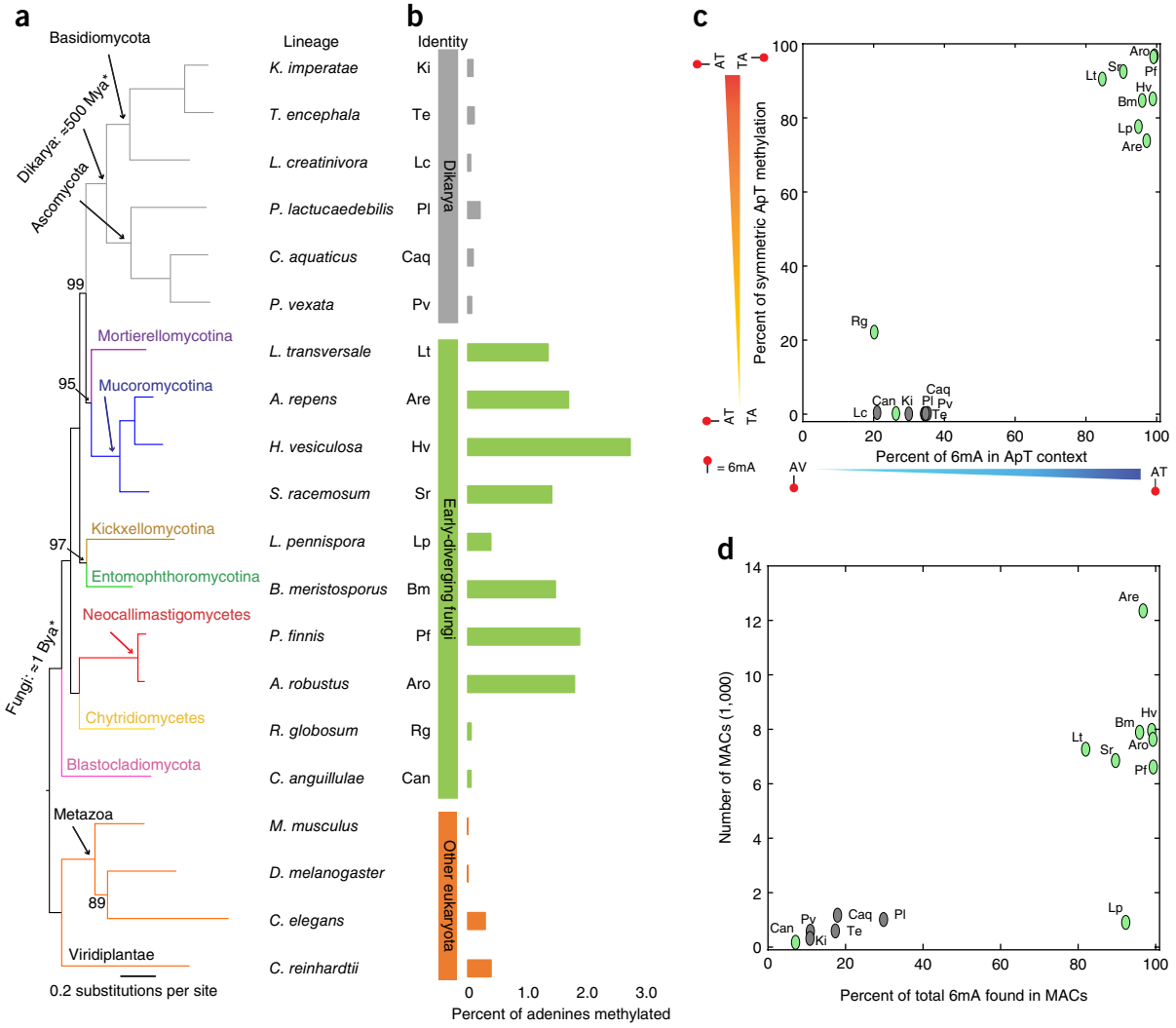

Figure 1 Phylogenetic diversity of genomes sequenced in this study and associated 6mA features. (**a**) RAxML (ref. 20) maximum-likelihood phylogeny constructed including 285 single-copy genes from all lineages sequenced in this study and four references in which 6mA has been analyzed previously[1–4]. Node ages indicate date of divergence, from ref. 5. Clades are colored at phylum, class and subphylum levels; bootstrap support for all nodes is 100 unless otherwise noted (shown in black at branches). Mya, million years ago; Bya, billion years ago. (**b**) Percent of total adenines methylated across all surveyed fungi plus outgroups. For additional data on cytosine methylation, see **Supplementary Figure 8**. ID abbreviations were generated for each lineage based on the first letter of the genus and first one or two letters of the species. (**c**) Methylation marks in early-diverging fungi are symmetric at ApT sites. The *x* axis shows the frequency of ApT methylation compared to any other adenine-containing dinucleotide (AV) combination. Although impacted by GC content, the frequency of ApT methylation by chance is expected to be around 25%. Colors and names of lineage groups as in **b**. (**d**) Percent of total 6mA marks found within MACs. No MACs were found in *R. globosum* or *L. creatinivora*.

(**Supplementary Fig. 2a**) and 6mA immunoprecipitation (IP) followed by sequencing (**Supplementary Fig. 2b,c**) for several lineages. While strong agreement across methods was observed for early-diverging fungi, additional validation experiments did not agree with SMRT-detected 6mA sites in the less-methylated Dikarya, suggesting that either (i) 6mA sites detected using SMRT analysis were largely false-positives in these genomes or (ii) due to their low abundance, it was difficult to accurately resolve 6mA presence and/or positioning using 6mA IP and mass spectrometry analysis (**Supplementary Note 1**). Methylation ratio analysis of individual 6mA sites showed that most positions are fully or nearly fully methylated (i.e., 100% of reads showed a methylation signature at a given position) in early-diverging fungi (but not Dikarya; **Supplementary Fig. 1d**), further supporting the notion that 6mA represents a reproducible and potentially functional epigenomic mark in these species.

Nearly all 6mA marks in early-diverging fungi were in the ApT context and symmetrically methylated (**Fig. 1c**). In 5mC eukaryotic DNA methylation, symmetric CpG methylation allows parental strands to carry marks through DNA replication as well as to propagate them to the newly synthesized child strand[13]. Given a comparable preference for symmetric methylation across these two epigenetic marks, it is plausible to expect that 6mA can also be propagated across nuclear division. In contrast, in Dikarya we found no preference for ApT sites and no symmetric methylation (**Fig. 1c**).

We found that 80–99.6% of 6mA marks in early-diverging fungi were concentrated in dense methylated adenine clusters (MACs; **Figs. 1d** and **2a–c**), while marks in Dikarya were largely unclustered. Despite slight variations of MAC characteristics (**Fig. 2c**), all early-diverging fungi displayed strong avoidance of 6mA deposition at one ApT-containing trinucleotide, TAT (and its reverse complement, ATA),

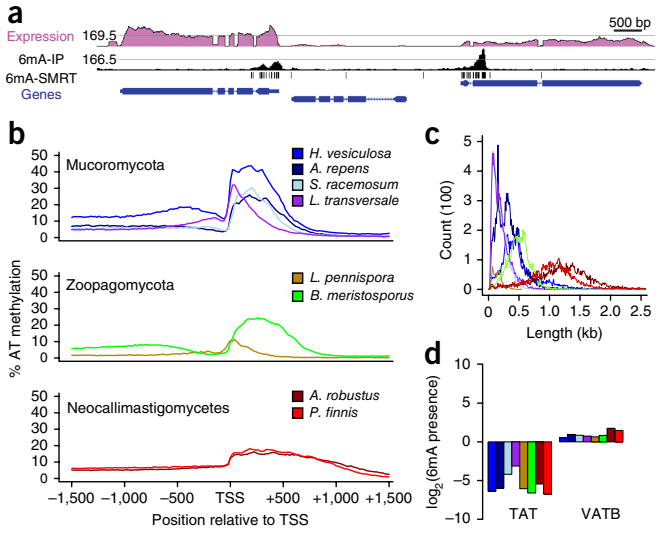

Figure 2 Distribution of 6mA marks across early-diverging fungal genomes. (**a**) Snapshot of 6mA deposition in *H. vesiculosa* (scaffold_1: 136200–146500). 6mA-IP, regions enriched in 6mA detected through 6mA-immunoprecipitation followed by sequencing; 6mA-SMRT; 6mA at single-base resolution detected through SMRT sequencing. Values shown next to expressions and 6mA-IP tracks represent the middle of their coverage ranges within this region. (**b**) Percent of ApT methylation surrounding transcriptional start sites (TSS) across early-diverging fungi harboring 6mA, separated by phylum and class. No enrichment for ApT methylation was found surrounding genes in *R. globosum*, *C. anguillulae* or any Dikarya. (**c**) MAC length distributions across all lineages harboring 6mA. While most MACs ranged between 100 and 750 bp, MACs found in Neocallimastigomycetes (red) were substantially larger (500–2,000 bp). (**d**) Prevalence of 6mA at TAT nucleotides versus VATB (IUPAC nomenclature for [C/A/G]AT[C/G/T]) sites. Colors in **c** and **d** are the same as in **b**.

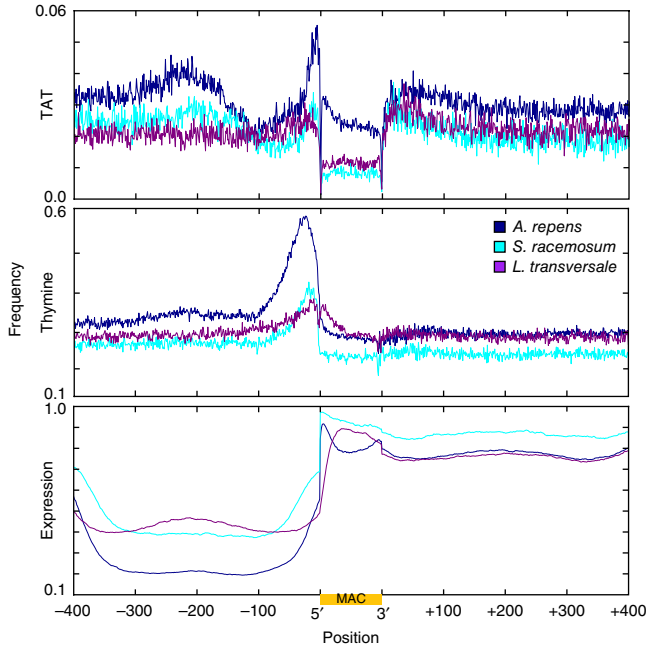

Figure 3 MAC characteristics across a subset of early-diverging fungi. For full set, see **Supplementary Figure 4**. Frequency of: TAT trinucleotides (top), thymine bases (middle) and expression (bottom) are plotted upstream, downstream and across MACs (orange bar). Frequency is calculated as number of occurrences ÷ total number of MACs. As MACs vary in length, all MACs ≥ 100 bp were selected, fragmented into 100 non-overlapping sections from start to end, and then average frequency was calculated within each fragment. MACs are oriented by gene direction.

which was almost never methylated (**Fig. 2d** and **Supplementary Fig. 3**). However, while we have identified several putative methyltransferases (**Supplementary Note 2** and **Supplementary Dataset 1**), the genes involved and mechanisms driving this avoidance remain elusive. We hypothesize that avoidance of TAT may be due to either a possible interference of 6mA with TATA-box functioning or steric hindrance impeding methyltransferase activity. However, methylation at other ApT-containing trinucleotides within MACs ranged from 49 to 92% (**Supplementary Fig. 3b**). These results demonstrate a highly sequence-context-dependent presence of 6mA in early-diverging fungi and support a possible role of TAT trinucleotides as functional components of MAC architecture.

Looking at the distribution of 6mA across early-diverging fungal genomes, we observed that almost all marks concentrated in promoter regions of protein-encoding genes, either at or slightly downstream of the transcriptional start sites (±500 bp; **Fig. 2a**,**b**). Furthermore, the structure and positioning of MACs appears to be influenced by nucleotide composition surrounding promoters. Notably, while we did not see enrichment of ApT dinucleotides within MACs, we did observe that TAT was depleted within MACs (higher frequencies of TAT within MACs was also associated with larger sizes; **Figs. 2c** and **3** and **Supplementary Fig. 4**) and that MACs often overlapped or were positioned immediately downstream of promoter thymine blocks (T-blocks; **Fig. 3** and **Supplementary Fig. 4**), shown previously in *C. elegans* to be involved in nucleosome-eviction[14]. In *C. reinhardtii*, 6mA marks spaced less than 150 bp apart (i.e., separated by less than a single nucleosome) lead to the absence of a nucleosome at that location[1].

Due to their proximity to T-blocks and their high density and large size, we hypothesize that 6mA in fungi also influences nucleosome–DNA interactions. With respect to gene transcription, we found that transcription initiation was often coincident with the start of MACs (**Fig. 3** and **Supplementary Fig. 4**).

6mA was strongly associated with gene expression. Typically about half of the genes in early-diverging fungi harbored MACs (**Fig. 4** and **Supplementary Figs. 5** and **6a**) and the vast majority of these were expressed (fragments per kilobase of transcript per million (FPKM) > 1.0)[1]. In contrast, unmethylated genes overall showed lower expression levels, and a much larger proportion of these lacked expression (FPKM ≤ 1.0; **Fig. 4** and **Supplementary Fig. 5**). Furthermore, through interrogating single-copy gene orthologs across early-diverging fungi, we found that 6mA presence was frequently conserved within ortholog clusters (83.6% of the 1,255 clusters investigated showed 6mA present across at least six of the seven lineages surveyed) and that when variability existed between lineages, 6mA presence significantly predicted expression (FPKM > 1.0) of orthologous genes ($P = 3.75 \times 10^{-13}$; hypergeometric test). Notably, while 6mA presence increased the likelihood of expression, the quantity of 6mA on genes did not have any clear impact on expression level (**Supplementary Fig. 5a**). These features suggest that while 6mA may act to facilitate transcription, the actual level of expression is regulated independently.

6mA was specifically tied to regulation of protein coding genes, as it was rarely found on predicted microRNAs[15] (**Supplementary Fig. 6b**) and almost never directly on tRNAs (**Supplementary Fig. 6c**). Whether this regulation occurs through cross-talk with translational machinery, RNA polymerase II or histone modification machinery (as in *C. elegans*[3]) is unclear. Highly conserved genes such as core

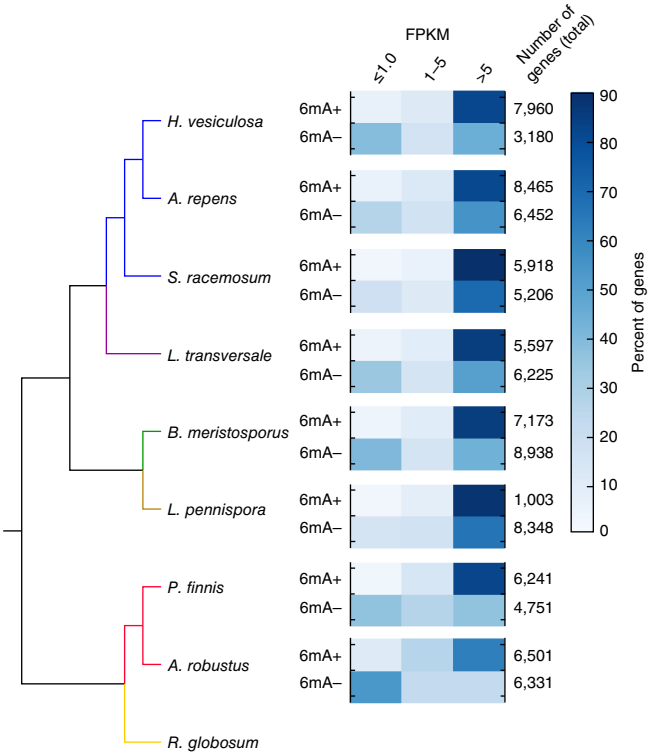

Figure 4 6mA is associated with active genes. Methylated genes rarely show expression below FPKM 1.0, whereas lack of expression is common in unmethylated genes. For each lineage, the percentage of methylated (6mA+) and unmethylated (6mA–) genes at a given FPKM level (≤ 1.0, 1–5, 5+) are shown, as well as total number of genes within each set. For a more complete breakdown of expression within methylated and unmethylated gene sets, see **Supplementary Figure 5**. Lineage colors are as in **Figure 1a**.

eukaryotic gene mapping approach (CEGMA[16]) and genes conserved across all fungi were the most frequently methylated (**Supplementary Table 2**), indicating that 6mA primarily targeted core genes in early-diverging fungi. Consistent with this, we found that MACs were not randomly distributed across genes in the genome. Rather, they were preferentially present or absent depending upon gene function (**Supplementary Fig. 7a**).

Although the presence of methylation at genes serving particular functions (assigned using Pfam)[17] occasionally varied by organism, many showed a consistent pattern across multiple lineages (**Supplementary Fig. 7b**). For example, we observed significant (false discovery rate (FDR) corrected $P \leq 0.05$, Fisher's exact test) over-abundance of 6mA across all lineages at several genes encoding constitutively expressed housekeeping proteins, such as mitochondrial Rho proteins (PF08477; involved in mitochondrial homeostasis), while some genes, such as those encoding leucine-rich repeat-containing proteins (PF00560; commonly involved in host–microbe interactions) showed a noteworthy lineage-specific variability in 6mA presence or absence. This suggests that 6mA is an important marker of expressed functionally relevant genes and that, while largely conserved, the types of genes methylated can also vary by lineage, perhaps in response to environment. For a full list of significant Pfam domains (FDR corrected $P \leq 0.05$, Fisher's exact test) for each lineage, see **Supplementary Dataset 2**.

The discovery 6mA in early-diverged fungi raises further questions regarding how it may interact with other epigenomic marks such as 5mC. To explore this, we investigated genome-wide 5mC abundance using bisulfite sequencing. We found that in genomes with high 6mA levels such as *S. racemosum*, *H. vesiculosa* and *L. transversale*, 5mC was nearly undetectable (**Fig. 1a**,**b** and **Supplementary Fig. 8a**,**b**). However, in *R. globosum* and *C. anguillulae*, where 6mA levels were negligible, we observed abundant 5mC (**Fig. 1a**,**b** and **Supplementary Fig. 8a**,**b**), suggesting a shift in epigenomic regulation in favor of 5mC in these chytrids. As expected, we also observed 5mC in Dikarya. These results indicate a strong negative relationship between the presence of one epigenomic mark and the presence of the other. Consistent with previous reports[6], in all genomes harboring 5mC methylation, it occurred primarily in the CpG context and was symmetrical (**Supplementary Fig. 8b**) and restricted to repeats (**Supplementary Fig. 8c**). In contrast, 6mA was enriched at promoters and uncommon at repeats (**Supplementary Fig. 8d**). Thus, 6mA and 5mC appeared to occupy distinct regions of the genome. Even in *Linderina pennispora*, the only genome that harbored both marks (at low levels; **Fig. 1a**,**b** and **Supplementary Fig. 8a**), we observed no overlap between marks.

Our results identify 6mA as a widespread epigenetic mark in early-diverging fungi associated with transcriptionally active genes. Given the age of the fungal kingdom[5,9] and the consistency of this signal within early-diverging fungi, it appears that for over 1 billion years the role of adenine methylation has been independently conserved. Notably, although fungi and animals are more closely related phylogenetically, we observed more similarity between 6mA profiles of early-diverging fungi and the alga *C. reinhardtii*[1] than between early-diverging fungi and Dikarya or any of the animals previously analyzed[2–4]. That these two distantly related kingdoms share common 6mA profiles raises the possibility that association of 6mA with gene expression is ancestral to the eukaryotic domain of life. Notably, it appears that over evolutionary time 6mA has evolved to serve two mutually exclusive roles: gene expression as in early-diverging fungi and algae[1]; and gene suppression, particularly of transposable elements, as reported in animals[2,4].

Alongside the evolution of dikaryotic mycelia, the drastic transition in 6mA utilization we observed represents a previously uncharacterized and likely critical difference separating Dikarya from other fungi. We suspect that the discovery of adenine methylation being strongly associated with gene expression may explain some of the historic difficulty in genetically modifying early-diverging fungi[18,19] and consequently may be crucial to developing successful transformation techniques within these (and other) challenging eukaryotic systems. For example, taking MAC positioning and features into consideration when designing constructs may be important for achieving methylation and sufficient expression of target genes. Our study illustrates how pronounced and widely used 6mA is as an epigenetic mark in eukaryotes. We anticipate that further characterization of 6mA will lead to fundamental transformations in our understanding of transcriptional regulation in this domain of life.

URLs. MycoCosm genome portal: http://jgi.doe.gov/fungi Links to individual portals include: *Hesseltinella vesiculosa* NRRL3301 (http://genome.jgi.doe.gov/Hesve2finisherSC), *Syncephalastrum racemosum* NRRL 2496 (http://genome.jgi.doe.gov/Synrac1), *Absidia repens* NRRL 1336 (http://genome.jgi.doe.gov/Absrep1), *Lobosporangium transversale* NRRL 3116 (http://genome.jgi.doe.gov/Lobtra1), *Linderina pennispora* ATCC 12442 (http://genome.jgi.doe.gov/Linpe1), *Basidiobolus meristosporus* CBS 931.73 (http://genome.jgi.doe.gov/Basme2finSC), *Piromyces finnis* (http://genome.jgi.doe.gov/Pirfi3), *Anaeromyces robustus* (http://genome.jgi.doe.gov/Anasp1), *Catenaria anguillulae* PL171 (http://genome.jgi.doe.gov/Catan2),

Rhizoclosmatium globosum JEL800 (http://genome.jgi.doe.gov/Rhihy1), *Clohesyomyces aquaticus* CBS 115471 (http://genome.jgi.doe.gov/Cloaq1), *Protomyces lactucaedebilis* 12-1054 (http://genome.jgi.doe.gov/Prola1), *Pseudomassariella vexata* CBS 129021 (http://genome.jgi.doe.gov/Pseve2), *Leucosporidiella creatinivora* 62-1032 (http://genome.jgi.doe.gov/Leucr1), *Kockovaella imperatae* (http://genome.jgi.doe.gov/Kocim1) and *Tremella encephala* 68-887.2 (http://genome.jgi.doe.gov/Treen1). Falcon assembler: https://github.com/PacificBiosciences/FALCON. BBDuk and other BBtools: https://sourceforge.net/projects/bbmap/. Biostrings: https://bioconductor.org/packages/release/bioc/html/Biostrings.html.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J.K. Henske, C. Swift, S.P. Gilmore and K.V. Solomon for preparing DNA and/or RNA for *P. finnis* and *A. robustus*; T. Porter for DNA and RNA preparation for *Catenaria anguillulae*; P. Liu for preparation of DNA and RNA for *R. globosum* and *L. transversale*; and D. Carter-House for preparation of genomic DNA of *H. vesiculosa* and *R. globosum* for bisulfite sequencing. For bisulfite sequencing of *H. vesiculosa* and *R. globosum*, we thank N.A. Rohr for library preparation and the Georgia Advanced Computing Resource Center (GACRC) for computational resources. Work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. This work was partially supported by funding from the National Science Foundation (DEB-1441715 to JES, DEB-1441604 to J.W.S. and DEB-1354625 to T.Y.J. and I.V.G.); Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work was further supported by the Office of Science (BER), US Department of Energy (DE-SC0010352) and the Institute for Collaborative Biotechnologies through grant W911NF-09- 0001. R.J.S. is supported by funding from the Office of the Vice President of Research at UGA as well as the Pew Charitable Trusts.

AUTHOR CONTRIBUTIONS

S.J.M., R.O.D. and I.V.G. designed the study. S.J.M. and R.O.D. collected and analyzed data under the supervision of G.K.-R. and I.V.G. R.C.K. optimized the protocol for 6mA IP-sequencing and PacBio library preparation. R.C.K. and C.D. sequenced genomes, including IP-sequencing. R.C.K., C.D., A.J.B. and R.J.S. conducted bisulfite sequencing. S.J.M., R.O.D. and A.J.B. analyzed bisulfite sequencing data. K.B.L., R.L. and T.R.N. conducted LC-mass spectrometry analysis. B.P.B. analyzed mass spectrometry data. K.L., B.B.A. and A.C. assembled genomes. S.J.M., S.H., A.K., S.R.A. and A.S. annotated genomes. A.L. and E.L. assembled transcriptomes. S.J.M., W.S. and G.K.-R. analyzed transcriptomes. A.G., D.C., J.M., T.Y.J., M.A.O'M., J.E.S., J.W.S. and I.V.G. coordinated genome projects. S.J.M. wrote the manuscript with significant input from A.V. and I.V.G.; and I.V.G. coordinated the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Fu, Y. *et al.* N6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. *Cell* **161**, 879–892 (2015).
2. Zhang, G. *et al.* N6-methyladenine DNA modification in *Drosophila*. *Cell* **161**, 893–906 (2015).
3. Greer, E.L. *et al.* DNA methylation on N6-adenine in *C. elegans*. *Cell* **161**, 868–878 (2015).
4. Wu, T.P. *et al.* DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* **532**, 329–333 (2016).
5. Lücking, R., Huhndorf, S., Pfister, D.H., Plata, E.R. & Lumbsch, H.T. Fungi evolved right on track. *Mycologia* **101**, 810–822 (2009).
6. Zemach, A., McDaniel, I.E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).
7. Blow, M.J. *et al.* The epigenomic landscape of prokaryotes. *PLoS Genet.* **12**, e1005854 (2016).
8. Fu, Y., Dominissini, D., Rechavi, G. & He, C. Gene expression regulation mediated through reversible m^6A RNA methylation. *Nat. Rev. Genet.* **15**, 293–306 (2014).
9. Taylor, J.W. & Berbee, M.L. Dating divergences in the fungal tree of life: review and new analyses. *Mycologia* **98**, 838–849 (2006).
10. Spatafora, J.W. *et al.* A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia* **108**, 1028–1046 (2016).
11. Zhang, W., Spector, T.D., Deloukas, P., Bell, J.T. & Engelhardt, B.E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.* **16**, 14 (2015).
12. Flusberg, B.A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
13. Breiling, A. & Lyko, F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics Chromatin* **8**, 24 (2015).
14. Grishkevich, V., Hashimshony, T. & Yanai, I. Core promoter T-blocks correlate with gene expression levels in *C. elegans*. *Genome Res.* **21**, 707–717 (2011).
15. Nawrocki, E.P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2015).
16. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
17. Finn, R.D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D1, D279–D285 (2016).
18. Obraztsova, I.N., Prados, N., Holzmann, K., Avalos, J. & Cerdá-Olmedo, E. Genetic damage following introduction of DNA in *Phycomyces*. *Fungal Genet. Biol.* **41**, 168–180 (2004).
19. Solomon, K.V. *et al.* Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. *Science* **351**, 1192–1195 (2016).
20. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

ONLINE METHODS

Strains and sequencing. To survey 6mA profiles over a broad range of fungi, we sequenced 16 fungal genomes using the PacBio sequencing platform. Lineages surveyed include 10 early-diverging fungi: *Hesseltinella vesiculosa* strain NRRL 3301 (Mucoromycota; Mucoromycotina), *Syncephalastrum racemosum* NRRL 2496 (Mucoromycota; Mucoromycotina), *Absidia repens* NRRL 1336 (Mucoromycota; Mucoromycotina), *Lobosporangium transversale* NRRL 3116 (Mucoromycota; Mortierellomycotina), *Linderina pennispora* ATCC 12442 (Zoopagomycota; Kickxellomycotina), *Basidiobolus meristosporus* CBS 931.73 (Zoopagomycota; Entomophthoromycotina), *Piromyces finnis* (Chytridiomycota; Neocallimastigomycetes), *Anaeromyces robustus* (Chytridiomycota; Neocallimastigomycetes), *Catenaria anguillulae* PL171 (Blastocladiomycota), *Rhizoclosmatium globosum* JEL800 (Chytridiomycota; Chytridiomycetes) and six Dikarya: *Clohesyomyces aquaticus* CBS 115471 (Ascomycota; Dothideomycetes), *Protomyces lactucaedebilis* 12-1054 (Ascomycota; Taphrinomycotina), *Pseudomassariella vexata* CBS 129021 (Ascomycota; Xylariales), *Leucosporidiella creatinivora* 62-1032 (Basidiomycota; Pucciniomycotina), *Kockovaella imperatae* (Basidiomycota; Agaricomycotina) and *Tremella encephala* 68-887.2 (Basidiomycota; Agaricomycotina). See URLs for links to individual MycoCosm genome portals.

For sequencing, fungal DNA was sheared to fragments > 10 kb long using a g-TUBE (Covaris). The sheared (or unsheared for *P. finnis*) DNA was treated with DNA damage-repair mix followed by end-repair and ligation of SMRT adapters using a PacBio SMRTbell Template Prep Kit (PacBio). For *L. pennispora*, *R. globosum* and *P. finnis*, DNA was then size-selected using a Sage Science BluePippen instrument. The prepared SMRTbell template libraries were sequenced on a Pacific Biosciences RSII sequencer using 2-h (*H. vesiculosa*), 3-h (*B. meristosporus*) or 4-h (all others) sequencing-movie runtimes. Genomes were assembled using Falcon (see URLs), improved with FinisherSC[21] (for some genomes) and polished with Quiver[22] (**Supplementary Table 1**). As it was designed for use on haploid genomes, we chose to run FinisherSC primarily based on ploidy statistics resulting from Falcon assembly. Each genome was then annotated using the JGI annotation pipeline[23]. Assembly methods and statistics as well as total 6mA bases identified are shown in **Supplementary Table 1**.

6mA modification detection and filtering. 6mA base modifications were detected with PacBio SMRT Analysis 2.3.0[12] using default parameters. Due to excessive repeats in *B. meristosporus*, it was not possible to complete SMRT analysis with enough coverage. Thus, we used RepeatScout[24] to mask repetitive sequences present >150× and used this as the input assembly for modification detection. 6mA was rarely found in repetitive sequences in other genomes; before hard-masking we performed a trial run using only five SMRT cells and found that the presence of 6mA in highly repetitive sequence was similarly negligible in *B. meristosporus*.

To control for potential false positives, after detecting kinetic signatures we removed any sites that fell below 15× per-strand coverage. Additionally, we calculated the upper limit of coverage calculated using the R boxplot function. Any value greater than the identified upper limit was removed to ensure that we could uniquely map each modification to an individual base, as well as reduce false 6mA discovery due to high coverage. Per-strand read coverage for all lineages and min/max cutoffs for detection of epigenetic modifications are shown in **Supplementary Figure 1a**,**b**. We further filtered results by selecting only 6mA marks with mQV ≥ 25 (**Supplementary Fig. 1c**).

Motif analysis and distribution of 6mA across the genome. In order to characterize the local sequence context of methylated adenines we used the Biostrings (wee URLs) and SeqLogo Bioconductor packages[25]. To interrogate the significance of the VATB (IUPAC for [G/A/C]AT[G/T/C]) motif, we surveyed all AT dinucleotides genome wide in each organism and calculated the frequency of all 16 tetramer sequence contexts (±1 nt from ApT); the ratios of tetramer abundance at symmetrically methylated ATs compared to the global distribution are plotted in **Supplementary Figure 3a**.

The Bioconductor package Genomic Features[26] was used to compare the overlap of gene annotations and 6mA modifications. Promoters were defined using the transcriptional start site of each gene, adding ± 500 nt of flanking sequence. Any remaining sequence from each gene was then classified as the gene body, and the rest of the genome was classified as either repeats[24] or intergenic. To visualize the distribution of methylation surrounding promoters, we gathered all AT dinucleotides relative to the TSS and calculated the percent methylation at each position across all genes (**Fig. 2b**).

LC-MS analysis of 6mA. Digested genomic DNA samples were subjected to mass spectrometry analysis based on methods previously reported[3,27]. Briefly, for sample digestion, nuclease-free water was added to 1–2 µg DNA for a final volume of 26 µL. DNA was denatured at 100 °C for 3 min, then chilled on ice for 2 min. DNA was then digested by adding 1 U DNase I (BioLabs) in 10 mM NH_4OAC buffer at pH 5.3 and incubating overnight at 42 °C. After overnight incubation, 3.4 µL of 1 M NH_4HCO_3 and 0.001 U phosphodiesterase I from *Crotalus adamanteus* venom (Sigma-Aldrich) were added and the sample was incubated at 37 °C for 2 h. Following incubation, 1 U alkaline phosphatase from *E. coli* (Sigma-Aldrich) was added and the sample was again incubated at 37 °C for 2 h. Finally, the digested DNA was diluted twofold with nuclease-free water and filtered through a 0.22-µm filter. In preparation for LC-MS analysis, 4 µL of an internal standard (2-amino-3-bromo-5-methylbenzoic acid (ABMBA), 10 µg/mL in 10% MeOH) was added to each sample of digested DNA (~68 µL total volume), then centrifuge-filtered through a 0.22-µm hydrophilic PVDF membrane (Millipore Ultrafree-MC). UHPLC reverse-phase chromatography was performed using an Agilent 1290 LC stack with a C18 column (Kinetex XB-C18, 1.8 µm, 100 A, 2.1 × 150 mm) at a flow rate of 0.5 mL/min with 1-, 2- and 8-µL injection volumes for each sample. The C18 column at 60 °C was equilibrated with 100% buffer A (100% H_2O with 0.1% formic acid) for 1 min, diluting buffer A down to 50% with buffer B (100% ACN with 0.1% formic acid) over 8 min, then up to 100% B over 1 min, followed by isocratic elution in 100% B for 1.5 min. MS and MS/MS data collection was performed using a Q Exactive Orbitrap MS (Thermo Scientific, San Jose, CA). Full MS spectra was collected from *m/z* 80–1,200 at 70,000 resolution, with MS/MS fragmentation data at 17,500 resolution using 10-, 20- and 30-V collision energies. dA and 6mdA were identified in samples based on *m/z* and on comparing retention time and MS/MS fragmentation spectra to purchased standards (N6-methyl-2′-deoxyadenosine, CAS 2002-35-9, Santa Cruz Biotechnology; 2′-deoxyadenosine monohydrate, CAS 16373-93-6, Sigma). Quantification of the 6mdA/(6mdA + dA) ratio was performed using calibration curves of the 6mdA and dA standards, each injected at volumes of 1, 2 and 8 µL, at concentrations ranging from 2.5 pg/mL to 100 µg/mL in water. Samples were measured for *S. racemosum* and *L. transversale* in digestion duplicates of 1–2 µg and for *H. vesiculosa*, *K. imperatae* and *C. anguillulae* on single samples of 1–3 µg.

6mA IP-seq. 6mA-IP was modified from the previous published 6mA RNA-IP protocol[28,29]. We sheared 3–5 µg of DNA to 200–400 bp using a Covaris E220 Ultrasonicator (Covaris). The fragments were treated with end-repair, A-tailing and ligation of Illumina compatible adapters (IDT, Inc.) using a KAPA-Illumina library creation kit (KAPA Biosystems). A portion of 10-µL ligated DNA was saved as an input control. We washed 50 µL of Protein A beads twice in 0.5 mL IP buffer (10 mM Tris-HCL (PH 7.4), 150 mM NaCl, 0.1% Igepal CA-630). We then preblocked 40 µL of the beads in 500 µL IP buffer containing 0.2 mg/mL of bovine serum albumin at 4 °C for 6 h. We used 10 µL of the beads to preclear the DNA at 4 °C for 2 h in 500 µL of IP buffer. The supernatant containing precleared DNA was then incubated with 0.5 to 1.0 µg of anti-6mA antibody (ABE572, Millipore) at 4 °C for 4 h. Then, the DNA–6mA-antibody mixture was incubated with the preblocked beads at 4 °C with gentle rotation overnight. The beads were washed four times with 1 mL of IP buffer and twice with 1 mL of Tris-EDTA (TE) buffer. The DNA fragments were enriched with 10–15 cycles of PCR and purified using SPRI beads (Beckman Coulter) for Illumina sequencing. The prepared library was then quantified using qPCR and run on the Illumina MiSeq sequencing platform, using a 1 × 50 run recipe. Following sequencing, input control and IP reads were aligned to each genome (*H. vesiculosa*, *S. racemosum*, *T. encephala* and *K. imperatae*) with BWA[30] using default parameters. Significant IP peaks (qval ≤ 0.01) were then called using MACS2 (ref. 31).

MAC identification and nucleotide composition surrounding MACs. Considering that 6mA is predominantly found in the ApT dinucleotide context

doi:10.1038/ng.3859

in early-diverging fungi, we calculated the distance between adjacent modifications and the number of ApTs traversed to find the ideal relative distance cutoff needed for optimal clustering of modifications. We scored clusters based on two criteria: the density (D) of 6mA at MACs (%6mA) and the efficiency (E) of clustering based on MAC size (MAC length/(MAC length + search distance)). We then calculated the sum of MAC scores (D × E) for all relative distances between 1 and 40. The search distance with the highest MAC score was used as our relative clustering distance to define MACs for each organism, which yielded a relatively normal distribution of MACs lengths (**Fig. 2c**).

We also explored distribution of various nucleotides surrounding MACs. For this, we counted occurrence of nucleotides within a flanking sequence (±400 bp) surrounding MACs, as well as across the MACs themselves (**Fig. 3** and **Supplementary Fig. 4**). As MACs vary in size, we fragmented each MAC into 100 non-overlapping pieces from start to end and then counted average nucleotide frequency within each fragment (i.e., observed nucleotide count ÷ total number of sites within fragment). The resulting counts were then normalized against the total number of MACs analyzed. For all distributions surrounding MACs, we only used MACs that directly overlapped either the TSS or the coding sequence (CDS) start site of a single gene to ensure proper orientation with gene direction. For thymine distribution, as MAC boundaries were always defined by methylation at ApT locations, artificial peaks and valleys in T frequency emerged at borders and were therefore trimmed.

Conservation of methylated genes. To survey 6mA presence with respect to gene conservation, we included these 16 genomes in a large ortholog clustering experiment (total: 196 taxa; **Supplementary Dataset 1**). We randomly sampled two published fungal genomes from MycoCosm[23] per family across all fungi and identified gene orthologs using mcl[32] (blastp[33]; cutoff: 1×10^{-5}; inflation factor of 2). For each genome, we determined gene conservation across the fungal kingdom at all taxonomic levels and then examined the frequency of methylation for each early-diverging lineage (**Supplementary Table 2**). Since we observed most MACs overlapped with genes slightly downstream of the TSS, we defined a gene as methylated if a MAC >100 bp was found ±500 bp from the CDS start. We also explored the methylation status of CEGMA genes (Core Eukaryotic Genes Mapping Approach[16]), which are expected to be present across all eukaryotes. CEGMA proteins were identified using BLAST with a minimum *e*-value cutoff of 1×10^{-5}.

Phylogeny reconstruction. For phylogeny reconstruction, we included all genomes analyzed in this study as well as four genomes previously described to harbor 6mA, including *C. reinhardtii*[34], *D. melanogaster* R6.09 (available on FlyBase[35]), *M. musculus* GRCm38.p5 (ref. 36) and *C. elegans*[37]. We then identified single-copy orthologs that were present in > 40% of taxa using mcl[32] (1×10^{-5}, inflation factor = 2) and aligned proteins with MAFFT[38] using default parameters. Ambiguous bases or uninformative sites were then removed using Gblocks[39] with the following parameters: –*t* = p, –e = .gb and –b4 = 5. This resulted in 285 single copy ortholog clusters, which were concatenated together for phylogeny reconstruction using RAxML[20] under the PROTGAMMAWAGF substitution model. We performed 100 bootstrap replicates.

Gene expression analysis. Except for *C. anguillulae*, for each lineage we generated stranded cDNA libraries using an Illumina Truseq Stranded RNA LT kit. mRNA was purified from 1 μg of total RNA using magnetic beads containing poly-T oligos. mRNA was fragmented and reversed transcribed using random hexamers and SSII (Invitrogen) followed by second-strand synthesis. The fragmented cDNA was treated with end-repair, A-tailing, adaptor ligation and 8 to 10 cycles of PCR. The prepared libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform using a TruSeq paired-end cluster kit, v3 or v2, and Illumina's cBot instrument to generate a clustered flowcell for sequencing. These flowcells were then sequenced on an Illumina HiSeq2000 or HiSeq 2500 sequencer using a TruSeq SBS sequencing kit, v3 or v4, for 200 cycles, following a 2 × 150 indexed run recipe.

For *C. anguillulae*, mRNA was purified twice from total RNA using an Absolutely mRNA™ purification kit (Stratagene). Subsequently, the mRNA samples were chemically fragmented to 200–250 bp using 1× fragmentation solution for 5 min at 70 °C (RNA Fragmentation Reagents, AM8740–Zn, Ambion). First-strand cDNA was synthesized using Superscript II Reverse Transcriptase (Invitrogen) and random hexamers. cDNA was purified with Ampure SPRI beads. Then the second strand was synthesized using a dNTP mix (with dTTP replaced with dUTP), *E. coli* RnaseH, DNA ligase and DNA polymerase I for nick translation. The dscDNA were purified and selected for fragments 200–300 bp using a double Ampure SPRI bead selection. The dscDNA fragments were then blunt-ended, poly-A-tailed and ligated with Truseq adaptors using an Illumina DNA Sample Prep Kit (Illumina). Adaptor-ligated DNA was purified using Ampure SPRI beads. Then the second strand was removed by AmpErase UNG (Applied Biosystems) similarly to the method described in Parkhomchuk *et al.*[40]. Digested cDNA was again cleaned with Ampure SPRI beads. Paired-end 76-bp reads were generated by sequencing using the Illumina HiSeq instrument.

Raw reads were filtered using the JGI QC pipeline. Briefly, this involved raw read evaluation for artifact sequence using BBDuk (see URLs) through k-mer matching (k-mer = 25), allowing 1 mismatch. Detected artifacts were trimmed from the 3′ end and RNA spike-in, PhiX and reads containing Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Finally, reads shorter than 25 bases or 1/3 of the original read-length (whichever was longer) were removed. FPKM values were calculated for each genome assembly using filtered RNA-seq reads mapped to gene models using Tophat[41] followed by Cufflinks version 2.1.1[42] using default parameters, with the exception of enabling fragment bias correction and reducing the maximum intron length to < 100,000 bp. Expression of each gene was then analyzed with respect to presence or absence of a MAC (±500 bp from CDS start; **Fig. 4**). Additionally, we identified bidirectional best-blast hits (BBH) across seven early-diverged fungi (*H. vesiculosa*, *A. repens*, *S. racemosum*, *L. transversale*, *B. meristosporus*, *A. robustus* and *P. finnis*) and then used single-copy BBH clusters (1,255 total), which varied in methylation status (626 clusters), to explore whether 6mA presence significantly predicted gene activity using the hypergeometric test (phyper function in R). For the purpose of genome annotation, transcriptome assemblies were built using RNNotator[43] for all genomes except those of *A. repens*, *C. aquaticus*, *P. vexata* and *T. encephala*, which were assembled using Trinity[44].

We also investigated expression directly upon MACs using filtered RNA-seq reads mapped to the genome using gmap[45]. Using the same approach as above (see "MAC identification and nucleotide composition surrounding MACs") we investigated how frequent expression was across the MAC, as well as ±400 bp flanking MACs (**Fig. 3** and **Supplementary Fig. 4**). MACs included in this analysis were a minimum of 100 bp in length, expression was normalized to the genome with the lowest number of RNA-seq bases mapped and a minimum cutoff of 10× read coverage was used to identify expressed bases. In addition to 6mA at genes, we found that MACs themselves typically marked regions of expression even in the absence of identified gene models. Since RNA was poly-A-selected before sequencing, these may represent candidate short peptides that traditional gene modeling fails to capture and may indicate 6mA as a potentially useful tool for tuning current gene modeling algorithms.

Separation of 6mA by gene function. To assess the functional consequences of 6mA presence, we explored methylation presence or absence at genes, separated by gene function (assigned using pfam version 30)[17]. For each genome, we analyzed genes harboring common pfam domains (present in at least eight genes) and compared methylation presence or absence to the expected frequency based on the total number of genes methylated in each genome (**Fig. 4**) using Fisher's exact test. Resulting *P* values were then FDR-corrected to adjust for multiple comparisons. Adjusted $P \leq 0.05$ were considered significant. See **Supplementary Dataset 2** for a complete list of results from Fisher's exact test.

Pfam domain analysis for methyltransferase identification. Pfam domains were identified using pfam version 30 (ref. 17) using the default search parameters. We included protein sequences from all taxa used in ortholog clustering (see above) and then searched these data for pfam domains that differed substantially in abundance between early-diverging fungi and Dikarya. Pfam counts across these two groups were compared using the independent two-sample *t* test followed by FDR correction. For a full list of pfam differences and

lineages included in ortholog clustering and pfam analysis, see **Supplementary Dataset 1**. BLASTp and tBLASTn (ref. 33) were used to search fungi for presence of previously discovered animal 6mA regulators DAMT-1, NMAD-1 and DMAD1 (refs. 2,3).

Bisulfite sequencing. For all genomes except *H. vesiculosa* and *R. globosum*, 1 µg of DNA was sheared to 500 bp using a Covaris LE220 (Covaris). DNA fragments smaller than 200 bp were removed using SPRI beads (Beckman Coulter). The fragments were treated with end-repair, A-tailing and ligation of methylated Illumina adapters (Illumina) using KAPA's Illumina library creation kit (KAPA Biosystems). The adaptor-ligated DNA was bisulfite-treated using EZ DNA Methylation Lightning Kit (Zymo Research), converting the nonmethylated nucleotides from cytosine to uracil. The converted DNA was enriched with ten cycles of PCR and then purified of PCR artifacts and size-selected using SPRI beads (Beckman Coulter). The prepared libraries were quantified using qPCR and run on the Illumina HiSeq sequencing platform, following a 2 × 151 indexed run recipe.

For the genomes of *H. vesiculosa* and *R. globosum*, single-end 150-bp MethylC-seq libraries were prepared according to Urich *et al.*[46] and sequenced on an Illumina NextSeq 500. Lambda-phage genomic DNA was used as a negative control to determine the efficiency of the sodium bisulfite conversion reaction.

Analysis of 5mC data. For the all genomes except *H. vesiculosa* and *R. globosum*, bisulfite-seq reads were quality controlled using BBDuk (see URLs), where they were first evaluated for artifact sequence by k-mer matching (k-mer = 23), allowing 1 mismatch. Detected artifacts were trimmed from the 3′ end of the reads; quality trimming was performed using phred trimming set at Q6; and reads 50 bp or shorter were removed. Finally, one base off the right end of each read was trimmed to prevent creation of completely contained read pairs. Reads were then mapped to each reference genome using Bismark version 0.16.3. For read mapping, bowtie1 (ref. 47) was used with seed length set to 40. Reads were mapped first in paired-end mode (max insert size set to 1,000), and unmapped reads were remapped individually using single-end mode. Prior to calling methylated cytosines, reads were deduplicated and reads that mapped to multiple locations were removed. Results were then combined for downstream analysis. Average per-strand cytosine coverage for each genome ranged from 11× (*S. racemosum*) to 44× (*C. anguillulae*), and bisulfite conversion efficiency ranged from 94.21% (*S. racemosum*) to 96.11% (*L. transversale*), based on spike-in controls.

For the genomes of *H. vesiculosa* and *R. globosum*, sequencing data was aligned to the genome using the methylpy pipeline[48]. Reads were trimmed of sequencing adapters using Cutadapt[49] and then mapped to both a converted forward strand (cytosine to thymine) and converted reverse strand (guanine to adenine) using bowtie[47]. Reads that mapped to multiple locations and clonal reads were removed. While the average per site coverage was 53.24 with an s.d. of 21.68 following mapping for *H. vesiculosa*, per-site coverage for *R. globosum* was very low (1.6×, s.d. = 1.47). Despite low depth, we still saw a strong 5mC signature from *R. globosum* at CpG dinucleotides and enrichment at repeats. Nonconversion rates were determined for both genomes using the lambda-phage control (0.16% for *H. vesiculosa* and 0.12% for *R. globosum*). Weighted DNA methylation was calculated for CG, CH (H = A|C|T) and all C sites by dividing the total number of aligned methylated reads by the total number of methylated plus unmethylated reads[50]. All 5mC results were then filtered by methylation ratio (minimum ratio = 0.2). With the exception of *R. globosum*, all libraries were also filtered by coverage (minimum 10× coverage).

Statistics. To explore whether 6mA presence significantly predicted gene activity, we used hypergeometric tests (phyper function in R); see "Gene expression analysis" above. For analysis of methylation presence or absence at genes based on their function, we analyzed genes harboring common pfam domains (present in at least eight genes) and compared the actual methylation presence or absence to the expected frequency based on total number of genes methylated in each genome (**Fig. 4**) using Fisher's exact test followed by FDR correction to adjust for multiple comparisons. Adjusted $P \leq 0.05$ were considered significant (see **Supplementary Dataset 2** for a complete list of results from Fisher's exact test). To identify putative methyltransferases involved in 6mA deposition, we explored pfam domains that differed significantly in presence or absence of methylation between early-diverging fungi (20 genomes) and Dikarya (176 genomes) using independent two-sample *t* tests (194 degrees of freedom) followed by FDR correction. Adjusted $P \leq 0.01$ were considered significant (see **Supplementary Dataset 1**).

Data availability statement. All data are available for each lineage through MycoCosm (see URLs) as well as deposited in GenBank under the following accession numbers: MCFA00000000 (*Clohesyomyces aquaticus* CBS 115471), MCFC00000000 (*Tremella encephala* 68-887.2), MCFD00000000 (*Linderina pennispora* ATCC 12442), MCFE00000000 (*Basidiobolus meristosporus* CBS 931.73), MCFF00000000 (*Lobosporangium transversale* NRRL 3116), MCFI00000000 (*Protomyces lactucaedebilis* 12-1054), MCFJ00000000 (*Pseudomassariella vexata* CBS 129021), MCGE00000000 (*Absidia repens* NRRL 1336), MCFL00000000 (*Catenaria anguillulae* PL171), MCGN00000000 (*Syncephalastrum racemosum* NRRL 2496), MCGO00000000 (*Rhizoclosmatium globosum* JEL800), MCGR00000000 (*Leucosporidium creatinivorum* 62-1032), MCGT00000000 (*Hesseltinella vesiculosa* NRRL 3301), MCFH00000000 (*Piromyces finnis*), MCFG00000000 (*Anaeromyces robustus*) and NBSH00000000 (*Kockovaella imperatae*).

21. Lam, K.-K., LaButti, K., Khalak, A. & Tse, D. FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics* **31**, 3207–3209 (2015).
22. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
23. Grigoriev, I.V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–D704 (2014).
24. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (Suppl. 1), i351–i358 (2005).
25. Bembom, O. *Sequence logos for DNA sequence alignments* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.431.3748 (2014).
26. Lawrence, M.. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
27. Yin, R. *et al.* Ascorbic acid enhances Tet-mediated 5-methylcytosine oxidation and promotes DNA demethylation in mammals. *J. Am. Chem. Soc.* **135**, 10396–10403 (2013).
28. Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201–206 (2012).
29. Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N. & Rechavi, G. Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing. *Nat. Protoc.* **8**, 176–189 (2013).
30. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
31. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
32. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
33. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
34. Merchant, S.S. *et al.* The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245–250 (2007).
35. Attrill, H. *et al.* FlyBase: establishing a Gene Group resource for *Drosophilamelanogaster*. *Nucleic Acids Res.* **44**, D786–D792 (2016).
36. Church, D.M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).
37. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
38. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
39. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
40. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
41. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
42. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
43. Martin, J. *et al.* Rnnotator: an automated *de novo* transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* **11**, 663 (2010).
44. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
45. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).

doi:10.1038/ng.3859

46. Urich, M.A., Nery, J.R., Lister, R., Schmitz, R.J. & Ecker, J.R. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10**, 475–483 (2015).
47. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
48. Schultz, M.D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–216 (2015).
49. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
50. Schultz, M.D., Schmitz, R.J. & Ecker, J.R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* **28**, 583–585 (2012).