# Supporting Information

## Chavez et al. 10.1073/pnas.1419513111

### SI Materials and Methods

**Computational Characterization of TET/JBP-Coding Transposons.** Iterative sequence profile searches were performed using the position-specific iterated (PSI)-BLAST and web version of the JACKHMMER (hmmer.janelia.org/search/jackhmmer) programs run against the nonredundant protein database of National Center for Biotechnology Information (NCBI). Multiple sequence alignments were built by the Kalign2 and Muscle programs, followed by manual adjustments on the basis of profile–profile and structural alignments. Similarity-based clustering for classification and culling of nearly identical sequences was performed by using the BLASTCLUST program (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html). The HHpred program was used for profile–profile comparisons. Secondary structures were predicted by using the JPred program. For previously known domains the Pfam database was used as a guide, although the profiles were augmented by addition of newly detected divergent members that were not detected by the original Pfam models. The TET/JBP genes were first identified by using the iterative sequence search procedures listed earlier. A custom Perl script was used to extract genomic context from the genome sequence of *Coprinopsis cinerea* from the NCBI RefSeq database (release 58) using a reference gene as the starting point (in this case the TET/JBP genes). The proteins encoded in these gene neighborhoods were then clustered by using BLASTCLUST with a bit-score density of 0.3 and Length parameter of 30% overlap. From the clusters thus obtained and the chromosomal coordinates of the corresponding genes, the extent of each Kyakuja transposon or its fragment was determined.

**Computational Analysis of Data from CMS-IP.** CMS-IP and CMS-input sequencing data were mapped against a fully in silico converted *C. cinerea* genome (*C.cinerea*_okayama7#130 including all supercontigs downloaded from the Broad institute) by using the Bismark software (1) version 0.6.4 (-q -n 2–chunkmbs 1028 bowtie-0.12.7). Mapping statistics for the technical and biological replicates are given in Table S3. Analogous mapping of lambda DNA against a fully in silico converted Lambda sequence revealed a conversion rate of 98.8% over all individual lanes of the CMS-IP and CMS-input replicates. CpG coverage, saturation, and correlation have been calculated by the Bioconductor software package MEDIPS (2) version 1.10.0 (extend = 250, uniq = T, window_size = 100, pattern="CG", BSgenome = custom genome generated from *C.cinerea*_okayama7#130). For the identification of 5hmC enriched regions (i.e., HERGs), we have first combined the mapping results of the technical replicates per biological replicate resulting in two independent CMS-IP and two independent CMS-input samples. Subsequently, we calculated differential coverage between the group of CMS-IP and the group of CMS-input samples by using MEDIPS version 1.10.0 (same parameters as earlier except window_size = 300, diff. method = edgeR, p.vale = 1e-3), resulting in 10,237 300 bp windows with significant differential coverage of which 9,736 were enriched in the CMS-IP samples over the CMS-input samples. Neighboring windows enriched in CMS-IP were merged (distance = 1) into 2,956 HERGs. In an alternative approach, we converted CMS-IP and CMS-input sequencing data (custom scripts) and mapped against the in silico converted *C. cinerea* genome using bowtie-0.12.7 (3) allowing for as many as 100 different mapping positions per read (-q -k 100–best–chunkmbs 512). Corresponding mapping statistics for the technical and biological replicates are given in Table S4. Subsequently, an alternative set of HERGs was calculated as described earlier (same parameter settings) but using the alternative multiple mapping results. Differential coverage analysis resulted in 10,640 significant 300-bp windows of which 10,576 were enriched in the CMS-IP samples over the CMS-input samples, which were merged into 2,142 HERGs. Finally, the set of 2,735 extended HERGs (2,623 located on the 13 chromosomes) was obtained by merging the two sets of HERGs (distance = 1) identified by allowing only unique or multiple mapping reads, respectively.

**Base Modification Analysis by Using SMRT Sequencing.** The sequenced data were mapped to the reference sequence and analyzed for base modification by using Pacific Biosciences' SMRTAnalysis (pacb.com/products/software/secondary-analysis) following a standard pipeline. The principle of base modification detection using SMRT sequencing by synthesis has been detailed in a white paper (pacb.com/pdf/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf). Briefly, the technique relies on the sensitivity of the polymerase kinetics to the DNA template structure as DNA synthesis is recorded in real time. The time between base incorporations, or IPD, is, on average, longer when nucleotide incorporation occurs opposite a methylated base in the DNA template compared with incorporation opposite a canonical base.

The analysis implements a *t* test for the sample IPDs against an in silico control at every position for identifying the modified sites. The alternate hypothesis in this test is that the sample set of IPDs stems from a population with larger IPDs than the in silico control, namely from incorporations opposite of a modified rather than canonical template base. For some visualizations (Figs. 3*E* and 4 and Figs. S2*E* and S3), and when calculating enrichments of modifications at genomic annotations, a threshold value of 60 for the log-transformed *P* value from the *t* test [called kinetic Qmod score, which is $-10\log(P \text{ value})$] at each reference position was used for flagging the given position as detected. This rather conservative threshold was applied to the native and digluc-SMRT sequencing data, and resulted in some modifications meeting this criteria in native SMRT but not in digluc-SMRT sequencing, likely because of slightly varying local coverages caused by differences in the SMRTbell template compositions of native SMRT (~6-kb insert libraries) and digluc-SMRT (from ~500-bp and ~6-kb insert libraries) sequencing.

For 5mC and its oxidized variants (5hmC, 5fC, and 5caC), the signature is not a single peak at the position of modification. Rather, there is a footprint of −6 to +3 bases relative to the modified site in the 5′–3′ direction (4). As the oxidation state of 5mC increases, the overall signal strength increases with little effect on the absolute footprint while, interestingly, maintaining a salient feature of having the strongest signals at positions 0, 2, and 6 bases in the 5′ direction from the modified site (Fig. 2*C*). Further, there is strong evidence of sequence context dependence on the relative signal intensities within the footprint signature. Although it is beyond the scope of this paper to discuss how one may harness the signature footprint to achieve base resolution for every sequence context around a modified 5C, it is important to note that these signatures are cumulative in nature when the modified sites are in close proximity (e.g., with overlapping signal footprint; Fig. S2*B*). This characteristic results in clusters of cumulated peak intensities for closely spaced modified cytosines and currently presents a challenge for deconvolution to obtain true base resolution in such regions. In the case of isolated 5mC and its variants, the signature may be properly assigned to the correct base. In such cases, one may then cluster all such sites as a function

of sequence context (MEME-ChIP) to identify potential target motifs.

**Bisulfite Methylation Analysis.** CMS-input sequencing data for all technical and biological replicates were pooled and mapped by using Bismark as described earlier, resulting in 3,984,911 million cytosines covered by at least one read (88.50% of all cytosines on both strands) and 2,851,987 million cytosines covered by at least 10 reads (63.34%). Methylation per cytosine was inferred by using the Bismark genome_methylation_bismark2bedGraph_v3.pl script.

**RNA-seq Data Analysis.** RNA-seq data of both replicates were mapped against the *C. cinerea* genome by using tophat (5) (-g 1 -r 115–mate-std-dev = 70–no-coverage-search–library-type = fr-secondstrand) using bowtie-0.12.7 (3). Read counts per gene (and other annotations) have been calculated based on the mapping results using custom scripts. To test, if zero read counts at several genes and other annotations is caused by exclusion of multiple mappers (due to -g 1), in an alternative approach, we also considered all hits of multiple mappers for counting reads per gene/annotation, but received similar results. Differential gene expression between oidia and hyphae has been calculated by using DESeq (6).

**Enrichment of HERGs and Digluc-SMRT Kinetic Variants in Known Annotations.** To calculate the significance of modifications (HERGs, digluc-SMRT kinetic variants) in known annotations, we performed random sampling of genomic regions: for each tested annotation, we generated 100 random sets of genomic regions, each set with the same number of genomic regions and with the same size distribution as the original annotation. For each random set, we counted the number of hits that fall into the random genomic regions, and modeled the resulting distribution by a normal distribution. Subsequently, we calculated the significance for the observed number of hits falling into the original annotations with respect to the modeled normal distribution derived by random genomic regions. Annotations marked by asterisks in Fig. 3 *B* and *C* and Fig. S6*D* are associated with a *P* value $\leq 0.001$.

**Genome Browser Visualizations.** All genome browser visualizations have been generated by using the Integrative Genomics Viewer (7).


1. Krueger F, Andrews SR (2011) Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27(11):1571–1572.
2. Chavez L, et al. (2010) Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res* 20(10):1441–1450.
3. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
4. Clark TA, et al. (2013) Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol* 11:4.
5. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105–1111.
6. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
7. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* 14(2):178–192.
8. Zolan ME, Heyler NK, Stassen NY (1994) Inheritance of chromosome-length polymorphisms in Coprinus cinereus. *Genetics* 137(1):87–94.
9. Stajich JE, et al. (2010) Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus). *Proc Natl Acad Sci USA* 107(26):11889–11894.

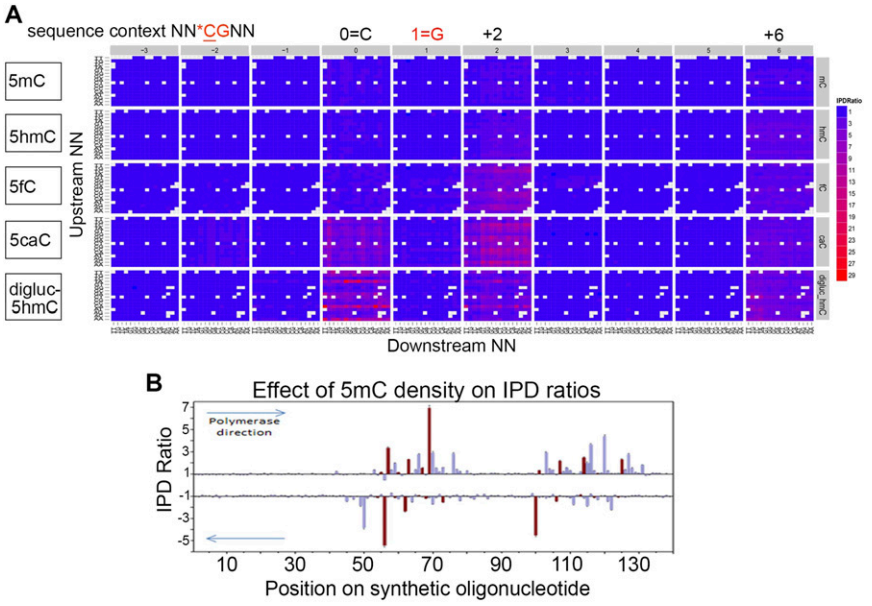
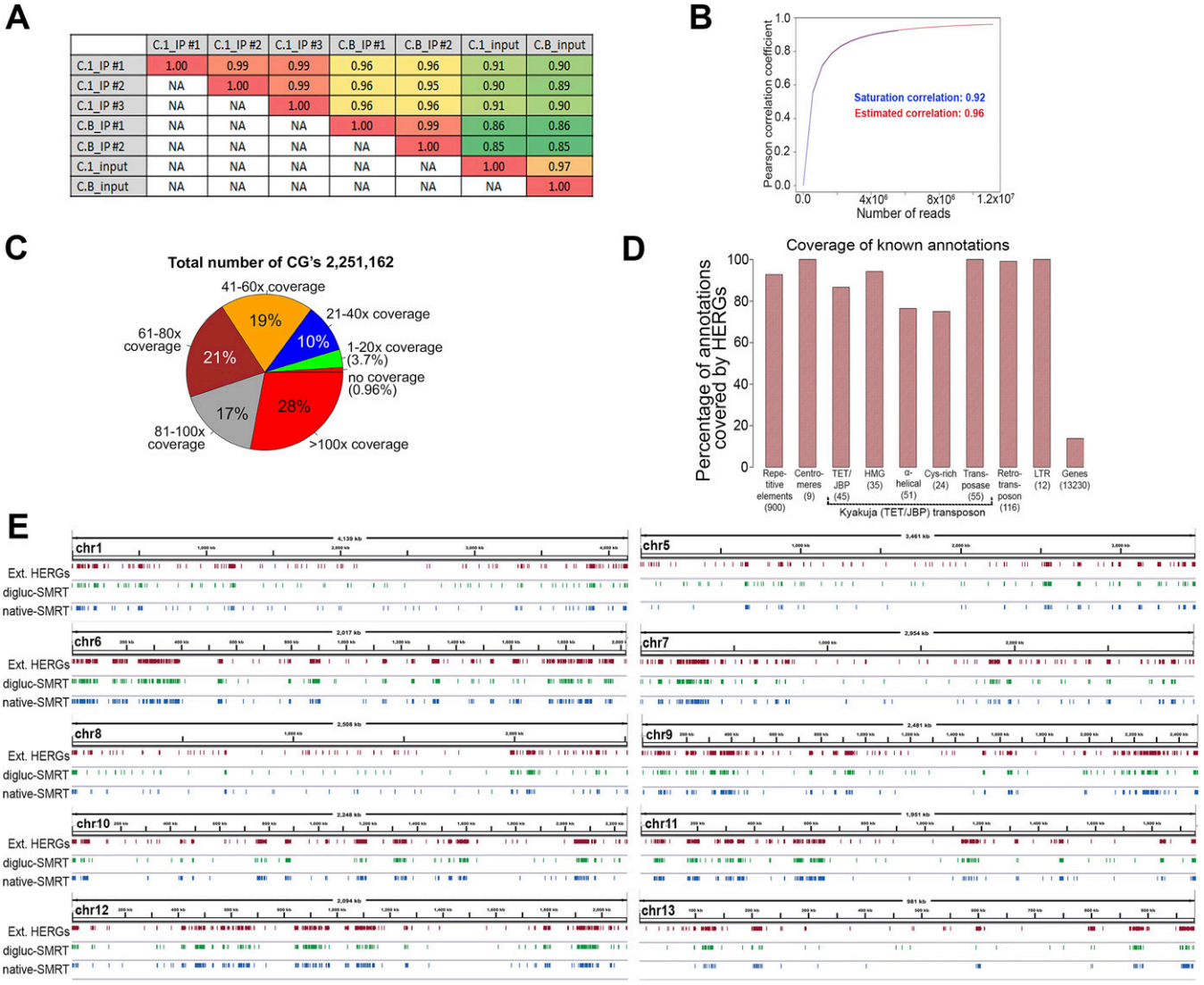




**Fig. S1.** (*A*) Heat map showing SMRT sequencing IPD ratio signatures for 5mC, 5hmC, diglucosylated 5hmC, 5fC, and 5caC at SMRTbell templates containing these modified cytosines in a randomized NNCGNN sequence context. (*B*) Effect of closely spaced 5mC. The IPD ratio kinetogram is for a synthetic oligonucleotide designed to have 5mC in close proximity, with red bars indicating the locations of 5mC. The effective signature is cumulative in nature: overlapping positions of 0, 2, and 6 in the 5′ direction of each the modified site show increasing signals that appear additive, albeit not necessarily in a linear fashion.




| | C.1_IP #1 | C.1_IP #2 | C.1_IP #3 | C.B_IP #1 | C.B_IP #2 | C.1_input | C.B_input |
|---|---|---|---|---|---|---|---|
| C.1_IP #1 | 1.00 | 0.99 | 0.99 | 0.96 | 0.96 | 0.91 | 0.90 |
| C.1_IP #2 | NA | 1.00 | 0.99 | 0.96 | 0.95 | 0.90 | 0.89 |
| C.1_IP #3 | NA | NA | 1.00 | 0.96 | 0.96 | 0.91 | 0.90 |
| C.B_IP #1 | NA | NA | NA | 1.00 | 0.99 | 0.86 | 0.86 |
| C.B_IP #2 | NA | NA | NA | NA | 1.00 | 0.85 | 0.85 |
| C.1_input | NA | NA | NA | NA | NA | 1.00 | 0.97 |
| C.B_input | NA | NA | NA | NA | NA | NA | 1.00 |

**Fig. S2.** (*A*) Pairwise Pearson correlations of short-read coverage at genome-wide 100-bp windows comparing the technical and biological replicates of the anti-CMS IP and input samples. Two biological replicates, C_1 and C_B, have three and two technical replicates, respectively. The correlation between technical replicates was 99%, and between biological replicates was >95%. (*B*) Saturation analysis (2) of the anti-CMS IP sequencing data shows sufficient saturation for a biological replicate with 11.2 million unique reads. (*C*) High CpG coverage of the ~2.25 million CpGs in the 13 assembled chromosomes of *C. cinerea* by the 11.2 million unique reads from one anti-CMS IP replicate. Reads were extended to a length of 250 bp along the sequencing direction corresponding to the estimated average fragment length of sonicated DNA. (*D*) Fraction of functional known and conserved genomic regions overlapping with extended HERGs. Almost all repetitive elements and Kyakuja (TET/JBP) transposons overlap with HERGs, indicating that they are strongly marked by oxi-mCs; in contrast, only a small fraction (13.8%) of genes overlap with HERGs. (*E*) Chromosome-wide views of all chromosomes except those already shown in Fig. 3*E*, illustrating the high agreement between the different techniques (anti-CMS, native SMRT, and digluc-SMRT sequencing).


**Fig. S3.** Additional examples of (*A*) a TET/JBP transposon and (*B*) a Dileera transposon, both embedded in a region with high abundance of oxi-mC. Units are non–strand-specific fractions (as percentages) of the sum of 5mC and 5hmC compared with all sequenced cytosines at each position deduced by bisulfite sequencing for mC oidia and mC mycelia, and strand-specific Qmod values for digluc-SMRT and native SMRT sequencing. (*C–E*) Additional examples of the distinct pattern of DNA modifications observed at 8 of 13 centromeres (Fig. 4*B*; units are as described earlier). (*C*) Example (chromosome 3) of a hypo-methylation-Copia-oxi-mC pattern present at four cytologically annotated centromeres, those of chromosomes 2, 3, 4, and 7. (*D*) The pattern was also present at three additional chromosomes 1, 6, and 12 (here shown for chromosomes 1 and 12). (*E*) The pattern on chromosome 10 occurred at a location other than the cytologically annotated centromere, which may reflect a strain difference: numerous chromosome-length polymorphisms have been described previously in *C. cinerea* (8).

| chr | position | strand | native ipdRatio | native coverage | native score | digluc ipdRatio | digluc coverage | digluc score |
|---|---|---|---|---|---|---|---|---|
| chr1 | 4131833 | 1 | 5.18 | 31 | 38 | 4.982 | 46 | 50 |
| chr1 | 4138863 | 1 | 3.138 | 57 | 47 | 5.077 | 53 | 87 |
| chr3 | 835055 | 1 | 3.078 | 44 | 62 | 5.43 | 47 | 61 |
| chr4 | 110264 | 0 | 3.698 | 38 | 41 | 6.773 | 58 | 64 |
| chr4 | 117279 | 0 | 4.193 | 30 | 35 | 4.49 | 36 | 49 |
| chr4 | 1121692 | 0 | 5.717 | 53 | 66 | 4.084 | 47 | 44 |
| chr4 | 1141444 | 0 | 4.316 | 54 | 58 | 6.754 | 37 | 55 |
| chr4 | 1540390 | 1 | 3.841 | 39 | 42 | 6.235 | 64 | 74 |
| chr6 | 245457 | 1 | 4.823 | 33 | 57 | 2.781 | 38 | 38 |
| chr6 | 1846581 | 1 | 3.916 | 33 | 39 | 5.675 | 57 | 58 |
| chr7 | 342079 | 0 | 2.169 | 36 | 30 | 5.126 | 49 | 68 |
| chr9 | 1062491 | 1 | 3.838 | 43 | 40 | 5.332 | 39 | 46 |
| chr9 | 2251632 | 1 | 3.688 | 43 | 38 | 5.405 | 46 | 58 |
| chr9 | 2258664 | 1 | 3.318 | 46 | 55 | 6.13 | 39 | 57 |
| chr11 | 1181574 | 1 | 4.309 | 32 | 47 | 4.34 | 42 | 40 |
| chr11 | 1189088 | 1 | 3.344 | 33 | 30 | 5.414 | 44 | 56 |
| chr12 | 1096651 | 0 | 4.058 | 80 | 90 | 4.552 | 70 | 71 |
| chr12 | 1103720 | 0 | 5.145 | 46 | 48 | 4.847 | 39 | 48 |
| chr13 | 968698 | 1 | 4.829 | 52 | 53 | 5.25 | 42 | 37 |

**Fig. S4.** Table showing the kinetic information (IPD ratios, coverage, Qmod/score) at single modified cytosines in the motif discovered in the LTRs of previously unknown retrotransposons, listed for native and diglucosylated samples.

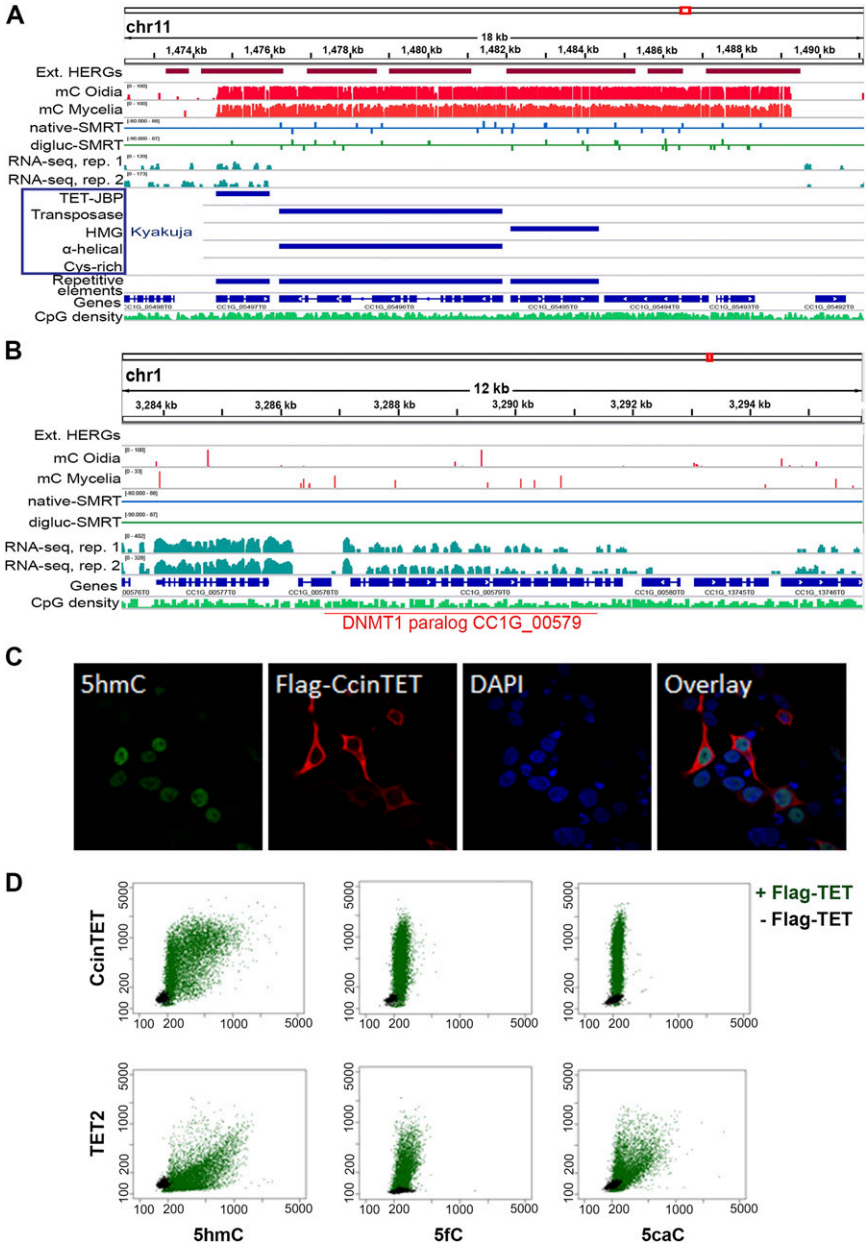



**Fig. S5.** (*A*) The single TET/JBP transposon, predicted to be catalytically active, that is moderately expressed at the oidial stage. Units are non–strand-specific fractions (as percentages) of the sum of 5mC and 5hmC compared with all sequenced cytosines at each position deduced by bisulfite sequencing for mC oidia and mC mycelia, and strand-specific Qmod values for digluc-SMRT and native SMRT sequencing. (*B*) One of the two DNMT1 paralogs that is moderately expressed in oidia and is free of any 5mC/oxi-mC (units as described earlier). (*C*) Immunostaining of FLAG-CcinTET (*C. cinerea* TET; *red*) and 5hmC (*green*) in HEK293T cells, depicting a microscope field different from that shown in Fig. 6*A*, in which the CcinTET protein is even more highly localized to the cytoplasm. We infer that the CcinTET protein shuttles between the nucleus and the cytoplasm, as we detect strong 5hmC signal in the nucleus of CcinTET-expressing cells. (*D*) Comparison of catalytic activity between CcinTET and human full-length TET2 using IXM high-content imaging system in HEK293T cells. Each dot indicates the intensity of DNA modifications (from left to right: 5hmC, 5fC, and 5caC) and CcinTET/hTET2 expression (anti-FLAG) in a single cell (*Top*, CcinTET; *Bottom*, hTET2). Green marks an HEK293T cell expressed with pEF-Flag-CcinTET or hTET2; black marks an HEK293T cell expressed with empty vector.

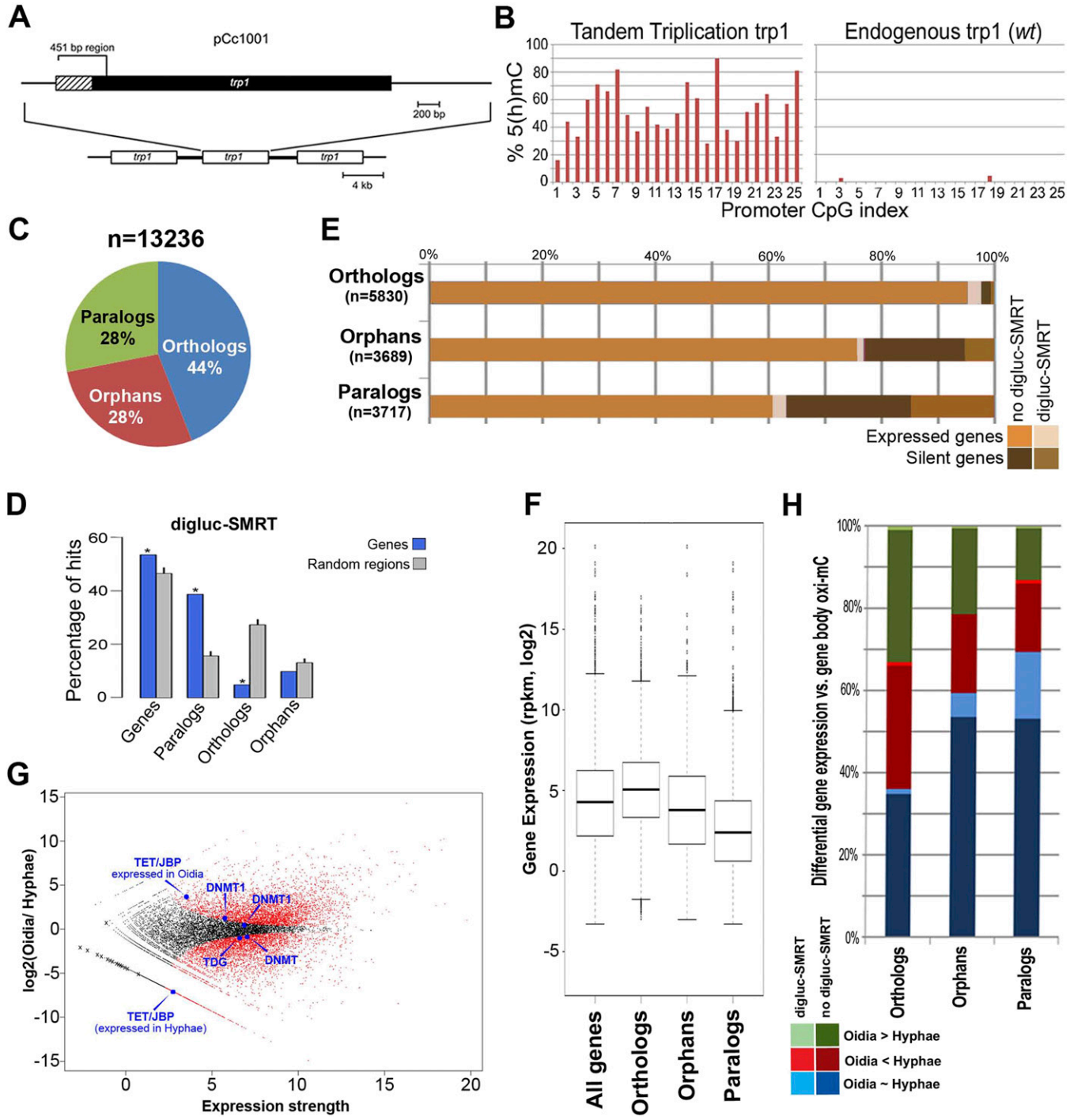



**Fig. S6.** (*A*) Diagram of the tandem triplication of the tryptophan synthetase gene. (*B*) The relative level of methylation and hydroxymethylation determined by bisulfite sequencing for each of the 25 CpG sequences within the 451-bp promoter of the tandem triplication (*Left*) compared with the endogenous trp1 promoter (*Right*). (*C*) The 13,236 protein-coding genes in the *C. cinerea* genome were previously classified into orphan genes with no obvious homologs in related species such as *Laccaria bicolor* (orphans, $n = 3,689$), single-copy genes with at least one ortholog (orthologs, $n = 5,830$), and paralogous multicopy genes that are primarily distributed in regions with average or high rates of meiotic recombination (paralogs, $n = 3,717$) (9). The pie chart shows the fraction of annotated orthologous, orphan, and paralogous genes in *C. cinerea*. (*D*) Shown are the fractions and enrichments of digluc-SMRT kinetic variants (Qmod ≥ 60; *Left*) at all genes, paralogs, orthologs, and orphans. Paralogous multicopy genes were more likely to be modified than single-copy genes, either orthologs which are significantly (empirical test, *Materials and Methods*) depleted of oxi-mC or orphan genes with no obvious homologs in a related species. (*E*) Fraction of expressed (beige) and silent (brown) genes subdivided into genes containing (light beige and light brown), or not containing (dark beige and dark brown), digluc-SMRT kinetic variants (Qmod ≥ 60). (*F*) Strength of gene expression for all genes after removing silent genes, separated into ortholog, orphan, and paralog genes. (*G*) MA-plot showing differential gene expression comparing oidia and hyphae. *x* axis: mean of library sized normalized RNA-seq counts in oidia and hyphae. *y* axis: log-twofold change between library size normalized RNA-seq counts in oidia compared with hyphae. (*H*) Fraction of genes with unchanged expression comparing oidia and hyphae (blue), increased expression in hyphae compared with oidia (red), and decreased expression in hyphae compared with oidia (green). These groups of genes are subdivided into genes containing (light blue, light red, and light green), or not containing (dark blue, dark red, and dark green), digluc-SMRT kinetic variants (Qmod ≥ 60).

**Table S1. Distribution of Kyakuja transposons and associated genes in *C. cinerea***

Table S1

**Table S2. Distribution of TET/JBP and Kyakuja transposase genes in *C. cinerea* chromosomes**

Table S2

**Table S3. Anti-CMS and input mapping statistics**

Table S3

Anti-CMS and Input mapping statistics for chromosomes 1–13 (top) and for the unassembled supercontigs (bottom).

**Table S4. Anti-CMS and input mapping statistics for chromosomes 1–13 and unassembled supercontigs when allowing for multiple alignments per sequence (maximum 100×)**

Table S4

Anti-CMS and Input mapping statistics for chromosomes 1–13 (top) and for the unassembled supercontigs (bottom) when allowing for multiple alignments per sequence (max. 100x).

**Table S5. Methylation and Hydroxymethylation (CMS input) in oidia**

Table S5

Statistics for genomewide C-to-T conversion rates (expressed at percentage of methylation/hydroxymethylation) at cytosines in CpG, CHG, and CHH context (H = A, T, or C) calculated based on bisulfite sequencing (CMS-Input) data. The top table shows the results for all covered cytosines and the bottom table shows the results for cytosines covered by at least ten sequencing reads.

**Dataset S1. RNA-seq read counts and differential expression analysis comparing oidia and hyphae**

Dataset S1