Supplementary Materials for Lea et al, *"Maximizing ecological and evolutionary insight from bisulfite sequencing data sets"*

Common pipeline for reanalysis of previously published data sets

Data processing

To investigate patterns of coverage and genome-wide DNA methylation in published bisulfite sequencing data sets (Table 1), we downloaded publicly available files from the NCBI Short Read Archive or contacted the authors to obtain files (details provided in Supplementary Table 1). In some cases, only FASTQ files were available, while in other cases text files providing the number of mapped methylated and unmethylated reads for each sample and measured CpG site were available. In all cases, we worked with the most processed file we could obtain.

For data sets where we obtained FASTQ files (Supplementary Table 1), we trimmed reads for adaptor contamination and base quality using the program 'Trim Galore!' (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Specifically, we trimmed bases with Phred scores below 20 and discarded any reads that were shorter than 20 base pairs long after trimming. We mapped the resulting trimmed and quality filtered reads to the following assemblies: (i) the *Cerapachys biroi* genome (*Cbir 1.0*) for data from[1]; and (ii) the domestic dog genome (*CanFam 3.1*) for data from[2]. We performed all read mapping using the bisulfite sequence aligner BSMAP[3]. For each aligned file, we extracted the number of methylated and unmethylated reads that mapped uniquely to a given CpG site using a Python script packaged with BSMAP. This step resulted in counts of mapped methylated and unmethylated reads for all samples in all data sets (obtained either through the above processing steps, by downloading text files directly from NCBI, or by contacting the authors). Finally, in each data set, we filtered for sites that were covered in >50% of all samples (Supplementary Table 1) and created data matrices describing (i) the number of methylated reads observed for each sample at each CpG site ('methylated counts matrix') and (ii) the total number of reads observed for each sample at each CpG site ('total counts matrix').

Estimating common properties of bisulfite sequencing data sets

For each data set, we used the processed total counts matrix to estimate the coverage properties presented in Supplementary Figure 1. Specifically, we first filtered for sites that had a median read depth (across samples) of >10x. For each data set, we then calculated the coefficient of variance of read depth, across all samples for each measured CpG site. This value provides information on variability in read depth across samples for each CpG site with at least 10x coverage.

We also focused on sites that had a median read depth (across samples) of >10x to estimate the mean and variance of DNA methylation levels in each data set. For each data set, we divided the filtered methylated counts matrix by the filtered total counts matrix to obtain estimates of DNA methylation levels that varied from 0 to 1. In each data set, we then calculated the mean and variance of DNA methylation levels on a site-by-site basis. We present the distributions of these values in Figure 3 and Supplementary Figure 5.

Estimating effect sizes

To estimate effect sizes for the predictor variables listed in Table 1, we performed further filtering on each data set. Specifically, for our effect size analyses we excluded sites that were hypermethylated or hypomethylated (mean methylation level >90% or <10%), invariant (sites that fell in the bottom 5% of the data set in terms of variance), or sequenced at low coverage (sites that fell in the bottom 25% of the data set in terms of mean coverage). We used the same filtering critera for each dataset with 1 exception: we relaxed the hypomethylation filter for the clonal raider ant data dataset (to exclude sites only with mean methylation levels = 0%) because the vast majority of the clonal raider ant genome is hypomethylated (Figure 3). Supplementary Table 1 reports the number of sites that passed these filtering criteria for each data set.

Next, we used the binomial mixed effects model implemented in MACAU[4] to test for an association between each predictor variable of interest and DNA methylation levels on a site-by-site basis. This analysis approach controls for familial relatedness or population structure among individuals in the data set by incorporating a random genetic effect determined by a user-defined pairwise relatedness matrix, *K*. For the clonal raider ant data set[1], the human cancer data set[5], and the human famine data set[6], we created *K* matrices based on the study design, which included samples from clones, repeated samples from the same individual, or full sibling pairs, respectively. A *K* matrix estimated from microsatellite data was already available for the yellow baboon data set[7]. For the *Arabidopsis* data set[8], we created a *K* matrix using SNP calls from publicly available whole genome sequencing data for the accessions included in our data set. Specifically, we filtered for sites called in at least 25% of samples, with a minor allele frequency ≥5%, and with a variant quality score ≥30. We then estimated a covariance matrix from this set of SNPs using the program GEMMA[9]. For the canid data set[2], where no genetic information was available, we called SNPs directly from the WGBS data using the program BisSNP[10] and used these genetic marker data to estimate the *K* matrix (see *Calling genetic variants from bisulfite sequencing data* for details). For the ape dataset, we used the percent sequence similarity estimates between species, as provided in[11] .

For each dataset, we ran MACAU for every CpG site that passed our filtering criteria using a model that included an intercept, the variable of interest as a fixed effect (Table 1), and a random effect that captured familial relatedness/population structure. We then extracted the beta estimate, β associated with the predictor variable of interest, x, for every analyzed site. Using these values, we calculated the proportion of variance explained by the predictor variable of interest, for each site, using the following equation:

$$\frac{\beta^2 * var(x)}{\sum \beta_j^2 * var(x_j) + \sigma^2}$$

Here, the denominator includes the beta estimate, β_j, for every fixed effect j, as well as the total sample variance, σ^2, which is estimated by MACAU. In all cases except the baboon data set, no additional fixed effects were included because information on relevant covariates was not available. In the baboon data set, we controlled for the age of the blood sample, the age of the individual at the time the sample was collected, and the bisulfite conversion rate as fixed effects[7]. Additionally, as an alternate measure of effect size, we calculated the mean difference in methylation levels between groups for data sets that included a binary predictor variable (Table 1). The distribution of estimates for both the mean difference between groups (for data sets with binary predictors) and the percent variance explained (the proportion from above multiplied by 100, which we abbreviate as PVE) are presented in Figure 3 and Supplementary Figure 2, respectively.

Reanalysis of previously published data sets

Estimating bisulfite conversion rates

We compared methods of estimating bisulfite conversion rate using an RRBS dataset from baboons[7]. First, using principal component analysis, we found that bisulfite conversion efficiency (estimated from a lambda phage spike-in) is associated with the first axis of variation in this data set. Second, we compared different approaches to estimating bisulfite conversion rate: spike-in, RRBS read ends (using cytosines introduced during the end repair step: Figure 1), and non-CpG methylation. We estimated bisulfite conversion rates from spike-in and non-CpG sites using the python script packaged with BSMAP. To estimate bisulfite conversion from the RRBS ends, we used Trim Galore! without the '-rrbs' option to retain the two bases on the 3'

ends of reads that ended with a CpG site. These two bases are added during the end repair step of the RRBS protocol and are therefore unmethylated. We mapped these reads to the baboon genome (using BSMAP) and calculated the proportion of times the 3' cytosine at an *Msp1* digest site was converted.

We simulated data to show the relationship between read depth and bisulfite conversion rate estimates using a binomial model. We varied the read depth between 1 and 250 reads and used simulated conversion efficiencies between 95 and 99.5%. Using the resulting data, we estimated the proportion of simulated samples (n=1000) for which the conversion efficiency was over/under-estimated and the mean absolute error in bisulfite conversion estimates from the data, relative to the original simulated conversion efficiency.

Calling genetic variants from bisulfite sequencing data (dog/wolf and Arabidopsis data sets)

For the dog/wolf data set[2], we called genotypes from BSMAP-mapped reads using BisSNP (Liu et al., 2012). Within BisSNP, genotype calls were filtered to retain only: (i) biallelic variants called in at least 25% of samples; (ii) variants with a minor allele frequency ≥5%; and (iii) variants with a variant quality score ≥30. For sample-specific genotype calls at these filtered variants, we considered only calls with a genotype quality score ≥20 based on at least two mapped reads. The resulting set of genotype calls were then used to generate a pairwise genetic covariance matrix using the program GEMMA[9].

To evaluate the use of BisSNP genotype calls for analyzing genetic effects (meQTL) and controlling for relatedness, we compared genotype calls made using BisSNP on 29 *Arabidopsis* accessions[8] to publicly available calls from whole genome resequencing data (from the 1001 Genomes Project: http://1001genomes.org/data/GMI-MPI/releases/v3.1/) and *Arabidopsis* array data available for 25 of the 29 individuals[12]. To estimate covariance matrices for these additional data sets, calls were filtered for biallelic sites with a minor allele frequency ≥5% and resequencing genotype calls were additionally thinned using *vcftools*[13] to include 1 variant per kb. Covariance matrices were compared with those derived from bisulfite converted DNA using a Mantel test.

meQTL analyses in the Arabidopsis data set

In the main text, we report effect sizes for analyses of local genetic variation on DNA methylation levels in *Arabidopsis* (i.e., *cis*-meQTL). To perform these analyses, we used publicly available SNP calls from the 1001 Genomes Project (http://1001genomes.org/data/GMI-MPI/releases/v3.1/) and DNA methylation data from[8]. Specifically, we took the set of filtered CpG sites from the DNA methylation data set (see *Estimating effect sizes*) and, for each CpG site, identified SNPs within 50 kb with ≥5% minor allele frequency, variant quality scores ≥30, and no missing genotype calls. Because this intersection and filtering criteria resulted in over 20 million possible tests, we randomly selected 1 million SNP-CpG pairs from chromosome 1 for our analyses. These results are reported in Figure 3 and Supplementary Figure 2.

Caste/phase effects and sample-specific methylation (clonal raider ant data set)

In the main text, we use a clonal raider ant data set[1] to illustrate a case in which the data are too low powered to detect an effect of interest (differences between reproductive and brood care phase ants) in site-by-site analyses, but show a globally apparent pattern. To assess power for site-by-site analyses, we simulated bisulfite sequencing data with the same read depth properties as the original data set, using the approach described in *Simulations*. In total, we created 615 simulated data sets containing 5000 sites each. Across data sets, we varied the proportion of true positive sites (i.e., the number of sites where β was not set to 0, which affects false discovery rate calculations) and the magnitude of the phase effect, β. We then converted the simulated count data to continuous methylation levels (m_{ij}/t_{ij}) and performed a paired t-test for each site, following the statistical approach in the original publication. Finally, we corrected

the distribution of 5000 p-values originating from each dataset for multiple hypothesis testing[14] and calculated the proportion of simulated true positives classified as significant at a 5% FDR cutoff.

For the global analysis on this data set, we used the *prcomp* function in R to perform principal components analysis on DNA methylation levels at all CpG sites that passed our initial filtering criteria. Specifically, for each of the sites passing filter, we converted DNA methylation levels to a normal distribution using the *scale* function in R, and then used linear models to control for line, sequencing batch, and read depth at each CpG site. We took the residuals from these models and used them to create a covariance matrix which served as the input for the PCA. After performing principal components analysis on the covariance matrix, we used a t-test to ask whether any of the top 5 PCs were significantly associated with caste. Finally, we permuted the caste labels of the samples 70 times (all possible combinations of 4 and 4), calculated whether any of the top 5 PCs in the permuted PCA were significantly associated with caste and, by counting the number of permutations that either a.) were significantly associated with a higher PC than the observed data, or b.) had a higher significant correlation value on the same PC as the observed data, calculated the significance for the observed data.

Finally, we used the same data set to investigate the practice of binarizing continuous data on DNA methylation, as in[1,15,16]. We suspected that this approach might generate "sample-specific methylation" even when the observed (continuous) methylation levels do not vary across samples. We extracted data for 100,000 random CpG sites that were sequenced to ≥10x across samples. For each sample and CpG site, we performed a binomial test and extracted the corresponding p-value. For each sample, we estimated the bisulfite conversion rate as the fraction of cytosines in a non-CpG context (CHH or CHG) that were converted to thymine (mean estimate ± s.d. = $0.992 \pm 6.28 \times 10^{-4}$). Finally, we corrected each sample-specific distribution of 100,000 p-values for multiple hypothesis testing[14]. Following the approach in Libbrecht et al., sites that passed a 5% FDR cutoff were considered methylated and all other sites were considered unmethylated.

Simulations

In silico *estimates of the properties of RRBS and WGBS data*

In Figure 1C, we report estimates of the proportion of cytosines in the human genome covered by an RRBS versus WGBS approach, as well as the distribution of these cytosines across different functional compartments of the genome. To obtain these estimates for an RRBS data set, we performed an *in silico* digest of the human genome (hg38) with *Msp1* (i.e., we 'cut' wherever the sequence CCGG was observed). We filtered the resulting fragments for sizes commonly retained during library preparation (100-500 bp), and then sampled 10 million fragments from this pool. Next, we mapped the first 100 bp and the last 100 bp of each fragment separately (using the default single-end mapping settings in the program BSMAP[3]) and retained all uniquely mapping reads. We then used *bedtools intersect*[17] to overlap the CpG sites contained within these reads with the locations of: (i) CpG islands (defined by the UCSC Genome Browser), (ii) CpG island shores (defined as the 2 kb regions flanking CpG islands), (iii) gene bodies (defined by RefSeq annotations), (iv) promoters (defined as the 2 kb region upstream of the TSS), and (v) regions far from genes (defined as regions >100kb from any annotated TES or TSS).

We performed a parallel analysis for WGBS, but instead of an *in silico* digest we performed *in silico* shearing of the human genome. To do so, we randomly chose 10 million locations in the human genome using *bedtools random*[17]. For each location, we extracted the sequence corresponding to a fragment of size *n,* with *n* drawn from a normal distribution with a mean of 300 bp and a standard deviation of 100 bp (thresholded to a range >100 bp and <500 bp). As described above, we mapped the first 100 bp and the last 100 bp of each of these

fragments, filtered for uniquely mapped reads, and tallied the number of CpG sites in our alignment file that fell in different functionally annotated genomic elements.

Impact of sample size and effect size on power to detect differential methylation

In the main text, we present results from several power simulations. To conduct these analyses, we first simulated a binary predictor variable of interest, x. The length of the x vector was equal to the simulated sample size, and in all cases this binary predictor variable was evenly distributed across classes. We next simulated DNA methylation levels, π, for each site j and each sample i, as a (logit-transformed) linear function of x and the effect size, β. In addition, we included an effect of random environmental variation (e), passed through a logit link. e_{ij} is drawn from a normal distribution with a mean of zero and variance of 1 (representing a moderate effect of random environmental noise).

$$log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = x_i\beta_j + e_{ij}$$

To translate methylation levels into count data (the observable data in bisulfite sequencing experiments), we simulated total read counts (t_{ij}) for each site j and each sample i from a negative binomial distribution.

$$t_{ij} \sim NB(r_j, p_j)$$

Here, r_j and p_j are negative binomial parameters estimated from real RRBS data[6]. Specifically, we generated sets of r and p parameters by fitting a negative binomial distribution to the total read count data from 100,000 randomly selected CpG sites in a real RRBS data[6]. To do so, we used the function 'fitdistr' in the R package MASS[18]. To simulate counts for a given CpG site, we randomly selected one of these parameter sets to produce the total number of reads. Finally, to simulate the number of methylated reads (m_{ij}) for each sample and each site, we drew from a binomial distribution parameterized by the resulting site- and individual-specific read depth (t_{ij}) and the DNA methylation level (π_{ij}):

$$m_{ij} \sim Bin(t_{ij}, \pi_{ij})$$

Using this simulation approach, we created datasets containing 5000 sites each. For all analyses presented in the main text, we set the proportion of true positive sites (i.e., the number of sites where $\beta \neq 0$) at 10%. In Supplementary Figure 3, we present results from identical simulations in which we varied the proportion of true positive sites. To assess power to detect the effect of interest in each simulated dataset, we ran a beta-binomial model on a site-by-site basis. Finally, we corrected the distribution of 5000 p-values originating from each dataset for multiple hypothesis testing[14] using the R function *qvalue*[19], and estimated the proportion of simulated true positives classified as significant at a 5% FDR cutoff.

Impact of variance in methylation levels on power to detect differential methylation

To simulate data sets with different variances in DNA methylation levels, we modified the procedures described above in the following ways. First, we projected the simulated DNA methylation level values (π_{ij}) for each site onto a normal distribution with a mean of 0.5 and a standard deviation value that we systematically varied. We did so using a quantile normalization approach, which allowed us to keep mean methylation levels constant, while exploring the effects of increasing variance. All other components of these simulations were performed as

described above. In total, we simulated data sets with four different levels of observed variance in DNA methylation levels (specifically, var(m_j / t_j) = 0.035, 0.045, 0.055, and 0.095; Figure 3C). We chose these values because they span common values observed in real RRBS and WGBS data sets (Supplementary Figure 5).

Binarizing sites as methylated or unmethylated

In the main text, we discuss the use of binomial tests[1,15,20] to classify individual sites as 'methylated' or 'unmethylated'. The typical use of this approach asks whether the number of methylated reads (m_{ij}) observed at site *j* in sample *i* can be explained by the rate of failure of the bisulfite conversion (i.e., 1 – the bisulfite conversion rate for sample *i*, b_i). If so, the site is classified as unmethylated, and if not, the site is considered 'methylated'. This binomial test is equivalent to evaluating the probability that m_{ij} originated from a binomial distribution defined by two parameters: the total read depth (t_{ij}) and the proportion of unmethylated cytosines in the sample that failed to convert to thymine ($1 - b_i$):

$$m_{ij} \sim Bin(t_{ij}, 1 - b_i)$$

This approach is influenced not only by the observed DNA methylation level at site j in sample *i* (m_{ij}/t_{ij}), but also by two technical factors: the total read depth at the site of interest (t_{ij}) and the bisulfite conversion rate (b_i). In addition, because the p-values from the binomial test are corrected using a false discovery rate approach[14], the genome-wide distribution of p-values (influenced by the parameters described above) will impact whether a given site is classified as 'methylated' or 'unmethylated'.

To understand how technical factors influence the categorization of sites as 'methylated' or 'unmethylated', we simulated bisulfite sequencing data spanning a range of observed DNA methylation levels and read depths. Specifically, we varied the read depth, t_{ij}, from 1 to 100, in intervals of 1 read. We also varied the observed DNA methylation level (m_{ij} / t_{ij}) from 0 to 0.3. We chose this range of DNA methylation levels because at levels > 0.3 all sites are considered methylated, regardless of coverage, at a nominal p-value cutoff of ≤ 0.05. Note that because m_{ij} must be an integer, the observed DNA methylation levels we were able to simulate for a given t_{ij} value vary. For example, for a total read count of 10 we could perform a binomial test for m_j = 0, 1, 2, or 3, equivalent to observed methylation levels of 0, 0.1, 0.2, or 0.3, respectively. However, for a total read count of 100, we could allow m_{ij} to take any integer value between 0 and 30, thus capturing a more granular set of possible observed methylation levels.

For each simulated combination of m_{ij} and t_{ij} values, we implemented the binomial test described above using the *binom.test* function in R[21] and set *b* equal to 0.992 (the average bisulfite conversion rate from[1] based on CHH and CHG conversion rates). Sites were classified as methylated if the observed number of methylated reads, m_{ij}, was unlikely to be observed by chance (i.e., if *p* < the chosen significance cutoff). Otherwise, the position was considered unmethylated.

The results of this analysis show that binarizing methylation levels using a binomial test leads to scenarios in which sites with exactly the same observed DNA methylation levels can be classified as either "methylated" or "unmethylated," depending on the total read depth and the significance alpha level (Supplementary Figure 6). For example, at a p-value cutoff of 0.01, a site with a DNA methylation level of 20% is called "unmethylated" if the total read depth is below 15x, but methylated if the read depth exceeds this coverage threshold. Furthermore, if the p-value cutoff is changed to $1x10^{-4}$, the same site is only considered methylated at read depths ≥20.

Comparison of approaches for detecting differentially methylated regions (DMRs)

In the main text, we report that the default run parameters for *BSmooth*, which focus on contiguous windows of 70 CpG sites (minimum), may not be directly translatable to RRBS data. In particular, we report that a very wide window (on average 34.474 kb) is needed to capture 70 CpG sites in human RRBS data, whereas much smaller windows containing 70 CpG sites are common in human WGBS data (average length = 2.938 kb in WGBS data). To obtain estimates of the window size needed to capture 70 CpGs, we focused on sites on chromosome 1 covered in >50% of samples in each of the two human data sets we reanalyzed (one RRBS[6] and one WGBS[5]). For each measured CpG site in each data set, we used a grid search approach to determine the start and end position of all possible windows around the focal CpG site that contained 69 additional CpG sites. For each CpG site, we retained the window with the minimum length. The numbers reported in the main text are summary statistics of the minimum window lengths across all CpG sites in the RRBS or WGBS data set, respectively.

To compare "site-first" versus "DMR-first" methods for identifying differentially methylated regions, we simulated bisulfite sequencing count data with correlational properties that mimicked a real RRBS data set. Specifically, we took all beta estimates for a binary predictor variable of interest from Lea et al.[7], focusing on analyzed CpG sites from chromosome 17 (n=12,562 CpG sites). Using these effect size estimates, we simulated methylated and total counts for each CpG site using the procedures described in *Impact of sample size and effect size on power to detect differential methylation* (where β_j was replaced with the beta estimates from real data for each site j). We simulated count data for a binary predictor variable across n=50 individuals, and we used coverage properties estimated from the same data set[7] as described above.

To test a "DMR-first" approach, we applied the program *BSmooth*[22] to the simulated RRBS data. Because RRBS data are less contiguous than WGBS data, we modified the default parameters and set the parameter 'ns', which corresponds to the minimum number of methylation loci in a smoothing window, equal to 8 instead of the default value of 70 because of the low density of CpG sites in RRBS data compared to WGBS data. *BSmooth*'s DMR detection algorithm searches for clusters of at least 3 differentially methylated sites (identified as sites with t-test statistic values that fall in the 5% most extreme set of test statistics observed in the data set) that fall within a 1 kb window; if two DMRs are <300 bp away from each other, they are combined.

To test a "site-first" approach on the same data set, we identified differentially methylated regions by first identifying differentially methylated sites using a beta binomial model, correcting the p-values for each site for multiple hypothesis testing[23], and then identifying the 2,141 sites in the simulated data set that passed a 5% FDR. For each of these 2,141 sites, we counted the number of nearby sites (within a 1 kb window centered on the focal CpG site) that also exhibited evidence for differential methylation (i.e., also passed a 5% FDR). We considered a region to be a DMR if at least 3 significant sites were found within a 1 kb cluster, and we combined any DMRs that overlapped by at least 1 bp.

With *BSmooth*, we identified 69 differentially methylated regions, compared to 55 identified with the beta-binomial approach. We note that as DMR size (i.e., the number of significantly differentially methylated sites) increases, there is increasingly greater overlap between the two sets of DMR calls (Supplementary Figure 9A). These results suggest that the most extreme DMRs are likely enriched for true positives, even though the total number and distribution of DMR sizes differs between data sets (Supplementary Figure 9B).

Supplementary Table 1. Data sets reanalyzed as part of this study.

Data set	Downloaded file type[1]	NCBI accession information	Sample size	Number of CpG sites used to estimate coverage properties[2]	Number of CpG sites used to estimate effect size[3]
Clonal raider ant	FASTQ	SRP066896	8	12,775,050	8,181,299
Yellow baboon	Text files	SRP058411	61	2,999,150	686,268
Human (famine)	Text files	GSE54983	48	2,565,213	716,764
Arabidopsis	Text files	SRP035593	30	33,150,775	2,630,875
Great apes	Text files (obtained from the authors)	SRP059313	4	1,301,272	540,604
Human (cancer)	Text files (obtained from the authors)	SRA036589	6	18,288,815	9,001,348
Dog and wolf	FASTQ	SRP065666	88[4]	10,511,958	363,326

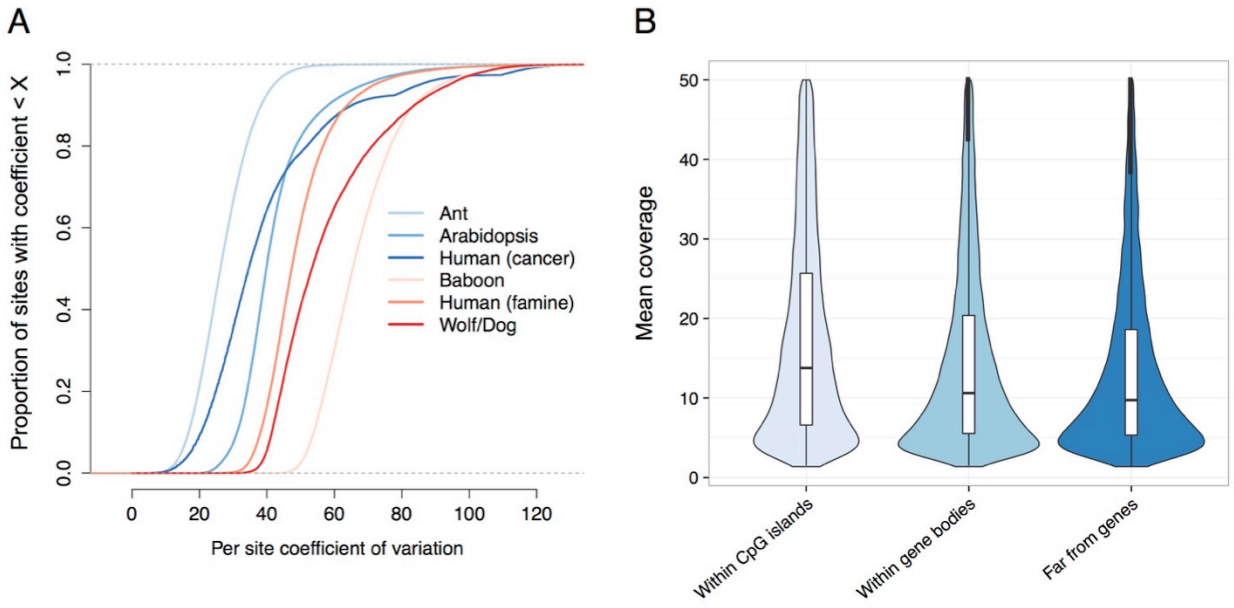
[1]For each data set, we either (i) downloaded raw FASTQ files from NCBI; (ii) downloaded text files denoting the counts of methylated and unmethylated reads, for each site and sample, from NCBI; or (iii) contacted the authors to obtain processed text files, when they were not available on NCBI.
[2]Number of CpG sites covered in >50% of samples in a given data set at a median read depth >10x
[3]Number of CpG sites covered in >50% of samples and further filtered for hypo and hypermethylation, variance in DNA methylation levels, and read depth (see Supplementary Materials). For large data sets (clonal raider ant, cancer, and Arabidopsis), we randomly sampled 1 million sites from the total set of filtered sites for effect size estimation.
[4]We excluded four samples from our analysis of this data set (2 dog samples and 2 wolf samples) because genotype data generated for these samples (see *Calling genetic variants from bisulfite sequencing data*) suggested that their FASTQ files were mislabeled. In particular, all four of the samples we excluded were labeled as one species (dog or wolf), but appeared genetically to cluster with the opposite species.

Supplementary Table 2. DMR (differentially methylated region) analysis programs.

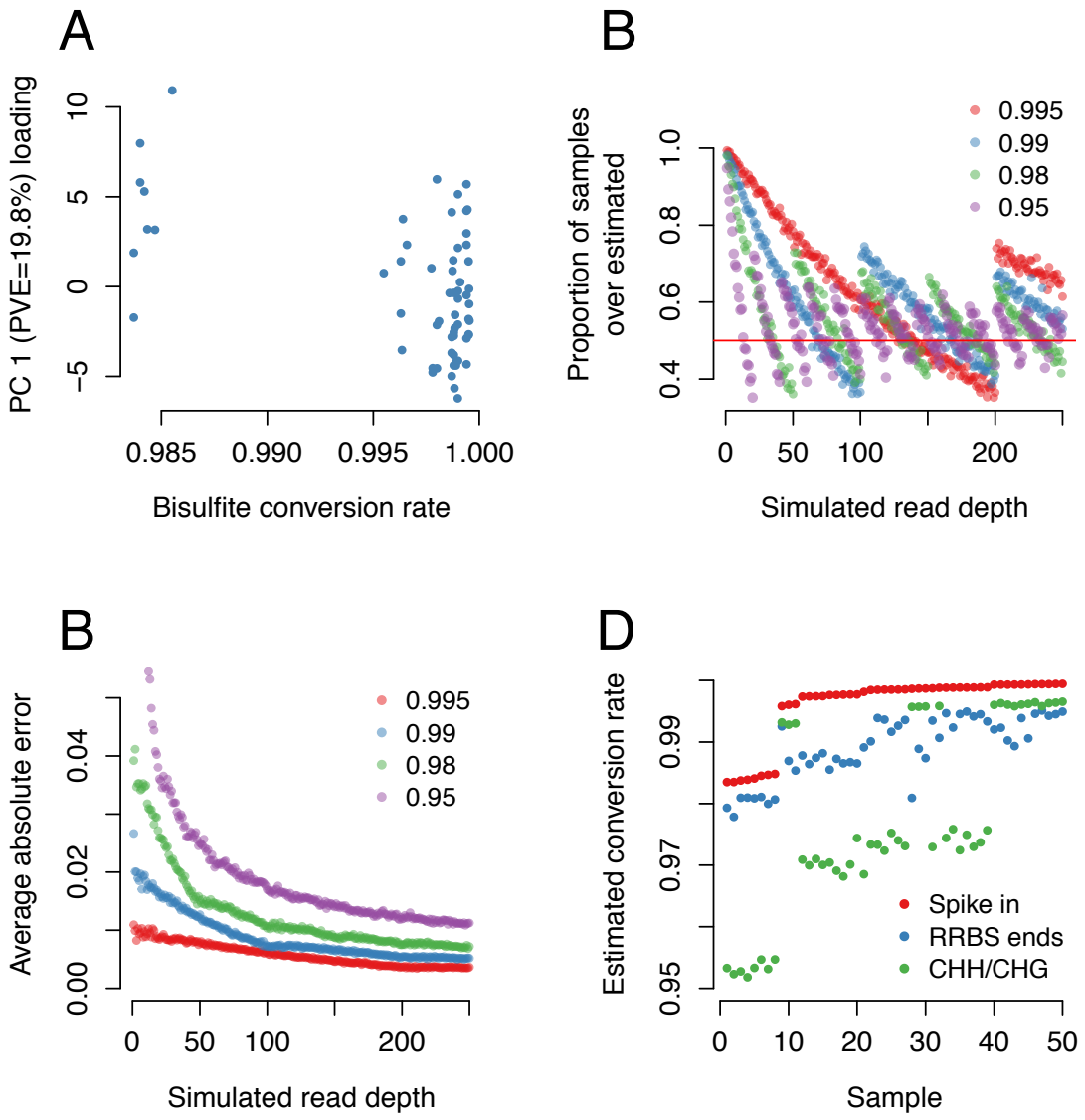
Program	Model	Smoothing[1]	Significance test	Accepts covariates?	Continuous predictor variable?
Bumphunter[24]	Linear regression	Yes	Permutation	Yes	No
BSmooth[25]	Linear regression	Yes	Signal-to-noise statistic	No	No
BiSeq[26]	Beta-binomial	Yes	Wald test	No	No
HMM-DM[27]	HMM	Yes	None	No	No
DSS[28]	Hierarchical beta-binomial	No	Wald test	No	No
RADmeth[29]	Beta-binomial	No	LRT, weighted Z test	Yes	No
MOABS[30]	Beta-binomial	No	Credible methylation difference	No	No
Metilene[31]	Circular binary segmentation	No	2D KS test	No	No
methylKit[32]	Logistic regression	No	Logistic regression	Yes	No[2]
eDMR[33]	Takes methylKit object as input	No	Stouffer-Liptak test	No	No

[1]'Smoothing' refers to methods that use local averaging of methylation levels or likelihood estimates to improve the precision of regional measurements and to borrow information across spatially proximate regions.
[2]Although none of the programs model continuous predictor variables, methylKit is able to accept multiple categorical predictor variables. All other programs test only for differences between two groups.

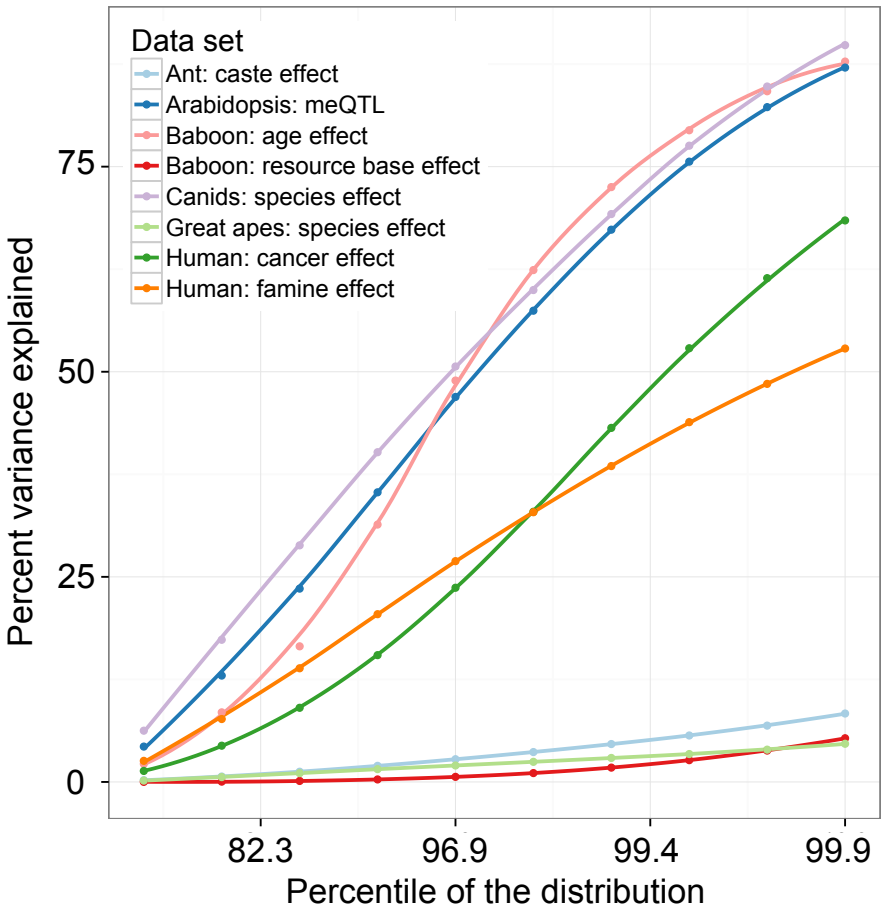
Supplementary Figure 1. Read depth variation in RRBS and WGBS data sets.
(A) For each data set, we calculated the coefficient of variation of read depth across all samples, for each measured CpG site with a median read depth of at least 10x. These values provide information on variability in read depth across samples for each CpG site. RRBS data sets are colored in shades of red while WGBS data sets are colored in shades of blue. Coverage in RRBS data sets is consistently more variable across samples. (B) Mean coverage of CpG sites measured via RRBS is consistently higher in CpG-dense regions compared to other genomic compartments. For every CpG site measured in >50% of the individuals in the yellow baboon data set[7], mean coverage for CpG sites that fell within the following three categories are plotted: (i) CpG islands, defined using UCSC Genome Browser annotations for the olive baboon genome, *Panu2.0*[34]; (ii) gene bodies, defined using Ensembl TES and TSS annotations for the olive baboon genome[35]; and (iii) regions far from genes, defined as regions >100kb from any TES or TSS. Whiskers on boxplots represent the values for the third and first quartiles, plus or minus 1.5x the interquartile range, respectively.



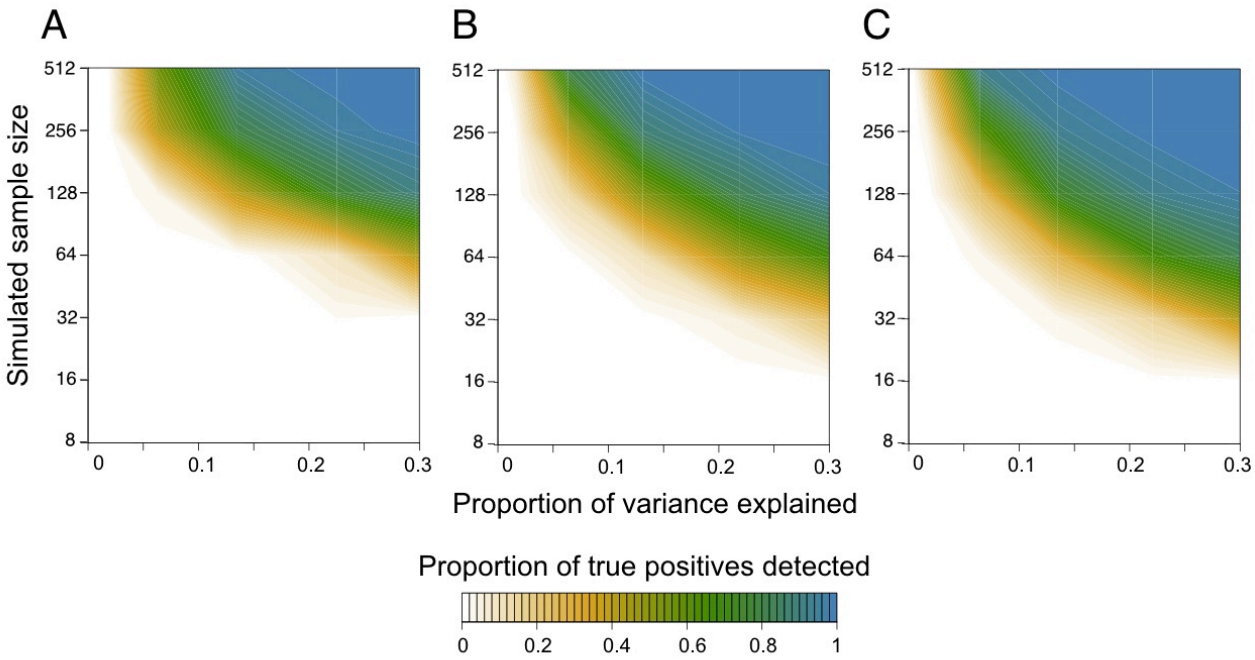
Supplementary Figure 2. Bisulfite conversion rate batch effects and estimation strategy.
(A) In a baboon data set[7], variation in bisulfite conversion efficiency produces a batch effect in which conversion rate estimates are correlated with PC1 of the overall sample-by-site data set (Spearman correlation, $p=4.91\times10^{-4}$, rho=0.433). (B) Probability of over- or underestimating bisulfite conversion rates in site-by-site analyses depends on read depth (x-axis) and the true bisulfite conversion rate (simulated values shown in different colors). (C) Expected error in site-by-site bisulfite conversion rate estimates, from simulated data (simulated values shown in different colors). (D) Comparisons between three strategies to estimate bisulfite conversion rate in the baboon samples[7]: based on a lambda phage DNA spike-in, end-repaired cytosines added in the RRBS protocol, and non-CpG (CHH or CHG) sites. All estimates are roughly correlated (Spearman correlation: spike-in vs. RRBS ends, $p<10^{-15}$, rho=0.804; spike-in vs. CHH/CHG, $p<10^{-15}$, rho=0.831; RRBS ends vs. CHH/CHG, $p=3.20\times10^{-6}$, rho=0.617).



Supplementary Figure 3. Effect size distributions for data sets reanalyzed here.
PVE values (y-axis) are plotted for select percentiles of the overall effect size distribution. To emphasize the extremes of the distribution, the x-axis is plotted on a log2 scale (ranging from the 50th to 99.9th percentile). Note that these results are also affected by other sources of variance in the data set. For example, PVE values for caste effects are low, which is partially explained by the strong effect of colony (median PVE>50%) in this data set. Further, both the cancer and great ape species data set have moderate PVEs, which can be explained by the extremely high levels of variability in DNA methylation levels in these data sets to begin with (Supplementary Figure 5).

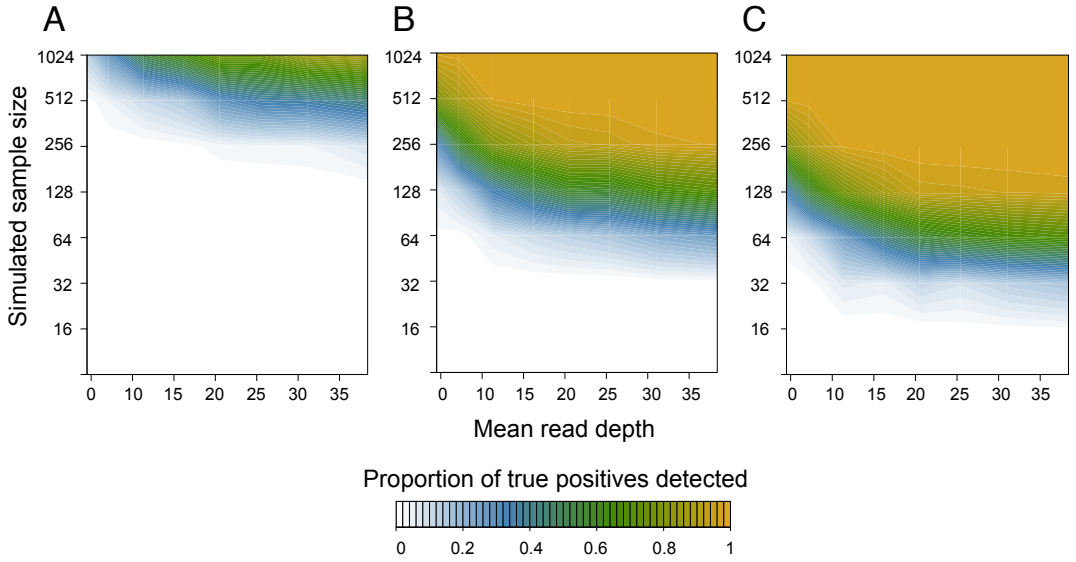


Supplementary Figure 4. Power to detect differential methylation.
We simulated data sets where the predictor variable of interest influenced methylation at (A) 1%, (B) 5%, or (C) 20% of all sites. Power to detect differentially methylated sites at a 5% FDR increases as a function of the simulated sample size, the magnitude of the effect of interest, and the proportion of simulated true positives. For example, to detect an effect that explains 15% of the variance with 50% power, a study would require a sample size of approximately 125, 90, or 65 depending on whether the proportion of true positives expected in the data set was 1%, 5%, or 20%, respectively.



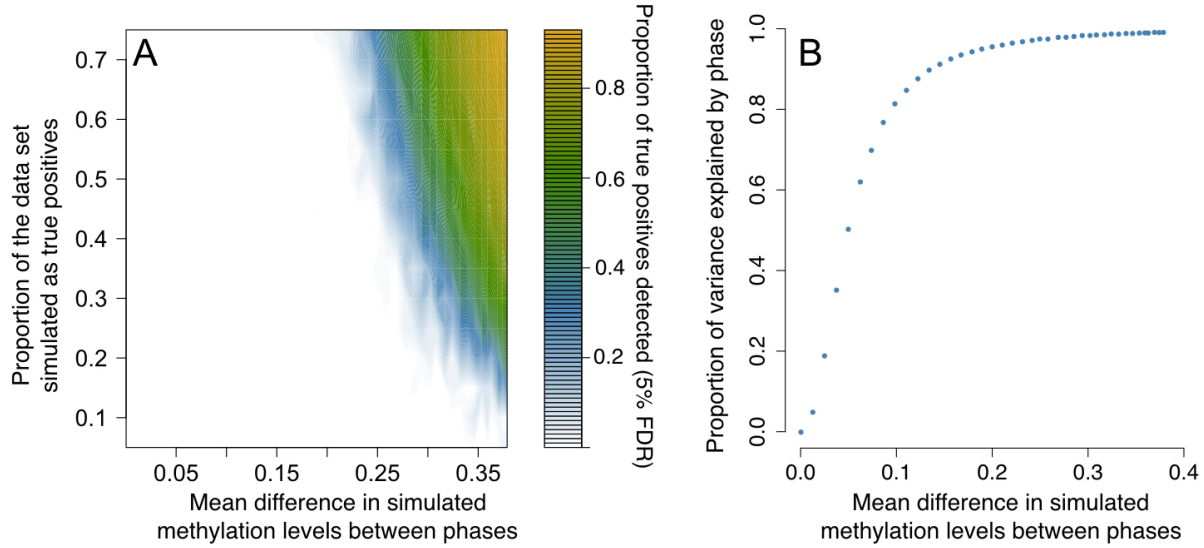
Supplementary Figure 5. Relationship between read depth, sample size, and power in simulated RRBS datasets.

We simulated datasets with small (5% mean difference between sample groups, panel A), moderate (10% mean difference between sample groups, panel B), and large (15% mean difference between sample groups, panel B) effect sizes, across a range of sample sizes and mean read depths, and calculated the proportion of differentially methylated sites detected at a 5% FDR. Note that sample size (y-axis) is plotted on a log scale. Across all effect sizes, increasing read depth beyond ~15-20x does not increase power (i.e., power does not increase as you move toward the extremes of the x-axis, for any given y-axis value); however, increasing sample size (i.e., moving up the y-axis) always increases power.



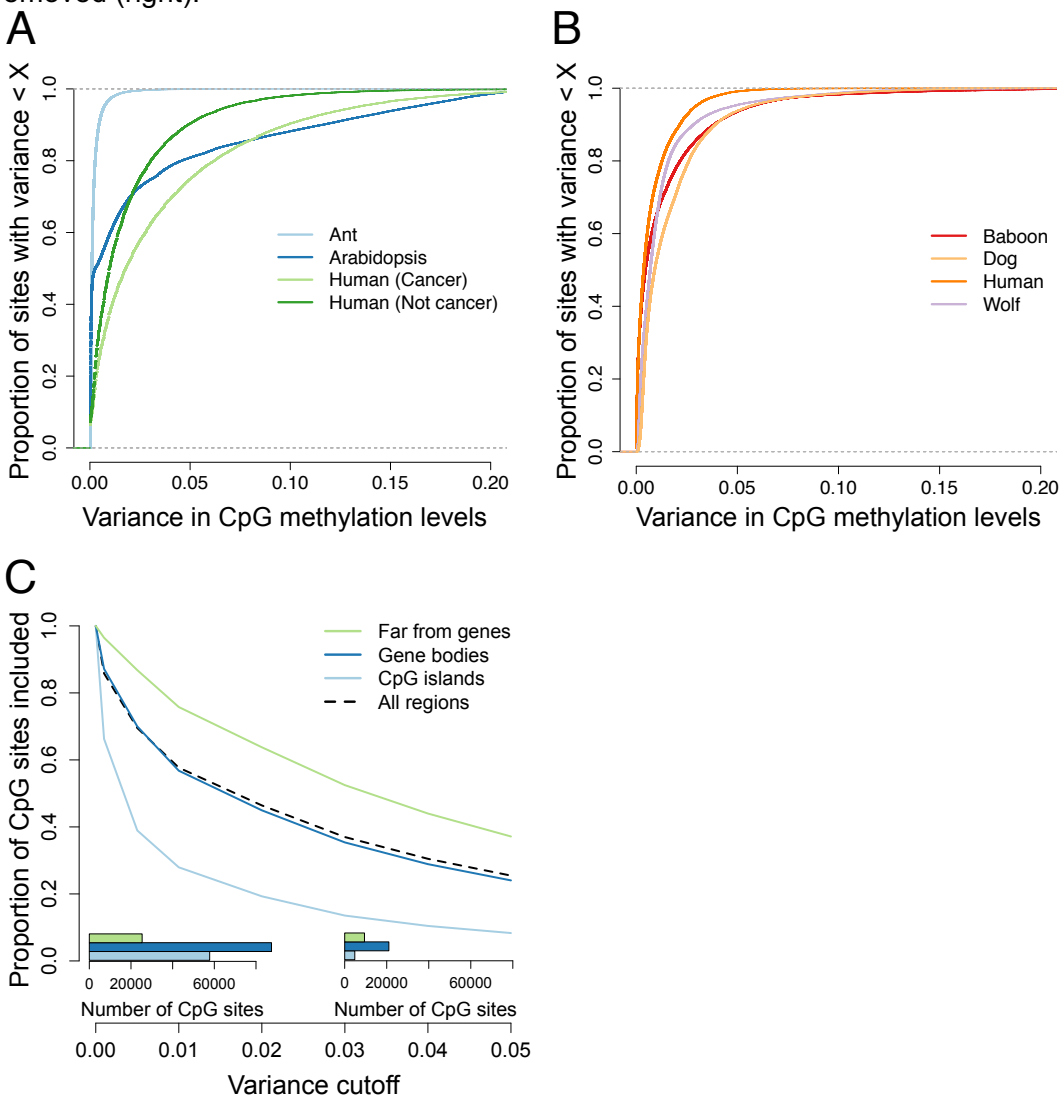
Supplementary Figure 6. Power to detect differential methylation between reproductive and brood care clonal raider ants.

Simulated data sets are based on the coverage properties of the WGBS clonal raider ant data set[1]. Across simulations, we varied the proportion of sites with a phase effect (i.e., the number of simulated true positives) as well as the effect size of the phase effect. (A) The shading shows the proportion of true positives detected at a 5% FDR, as a function of the mean difference in methylation levels between phases (x-axis) and the proportion of total sites in the data set that were simulated to have a phase effect (y-axis). These results show it is nearly impossible to detect differential methylation between phases unless phase is the primary source of variance in DNA methylation levels for many sites. (B) Comparison between two measures of effect size (mean difference in DNA methylation levels versus proportion of variance explained) for the simulated phase effect. Note that once the simulated difference between phases exceeds ~10%, almost all of the variation in DNA methylation levels is explained by phase.

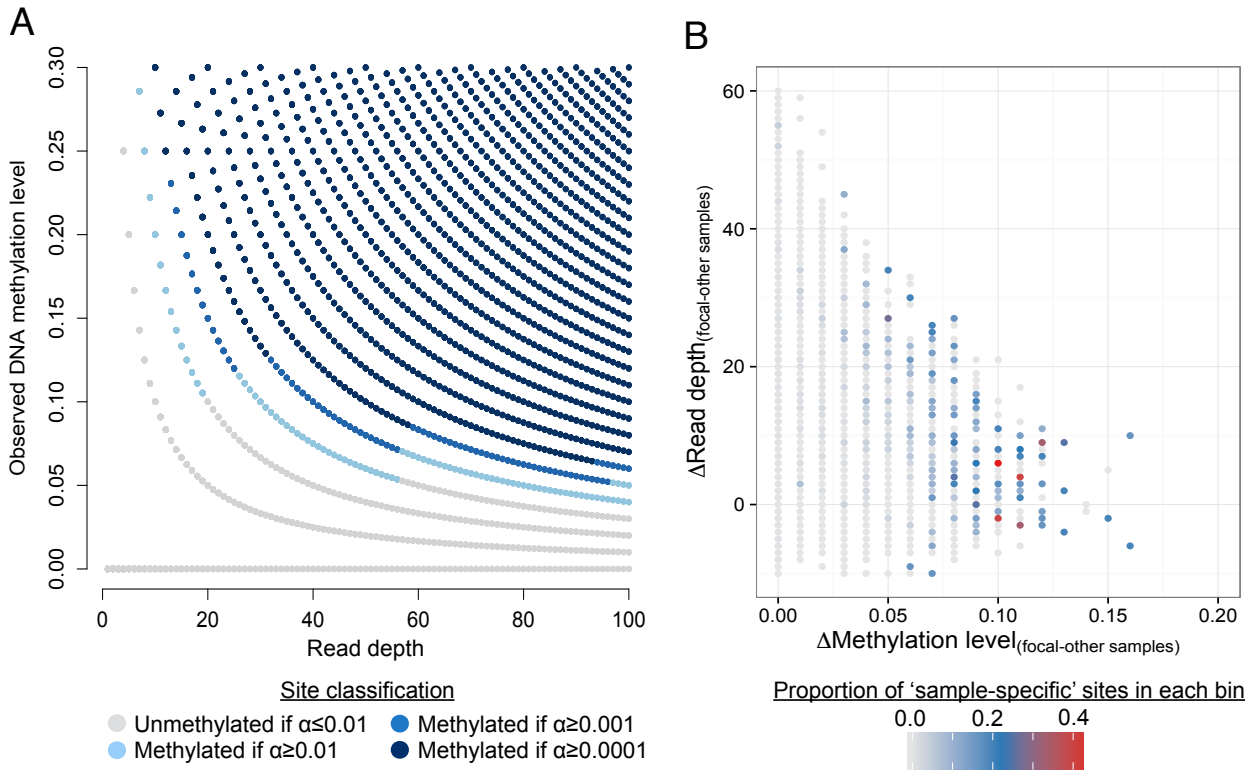


Supplementary Figure 7. Variance in CpG methylation levels across data sets.

(A-B) For each data set, we calculated the variance of DNA methylation levels at each CpG site with a median coverage >10x across all samples in the study. The distribution of variance estimates is presented as a cumulative distribution plot, where the y-axis represents the proportion of sites in each dataset with an estimated variance less than the x-axis value. (C) Removing low variance sites from an RRBS data set biases the set of analyzed sites towards regions of the genome that are intrinsically more variable. We calculated the variance in DNA methylation levels for all sites in the baboon RRBS data set[7], and then systematically filtered the data according to the variance cutoff shown on the x-axis. For each filtered data set (with variance greater than or equal to the x-axis cutoff), the proportion of CpG sites retained in the filtered data set relative to the original unfiltered data set is shown. CpG islands, which tend to be invariant and hypomethylated, are removed at a faster rate than more variable regions such as gene bodies and regions far from genes. Thus, removing sites with low variance (a filtering step that can increase overall power) will tend to alter the types of sites in a data set. Insets show the number of CpG sites in gene bodies, CpG islands, and regions far from genes in a data set with no variance filtering (left) versus a data set where sites with variance < 0.05 are removed (right).

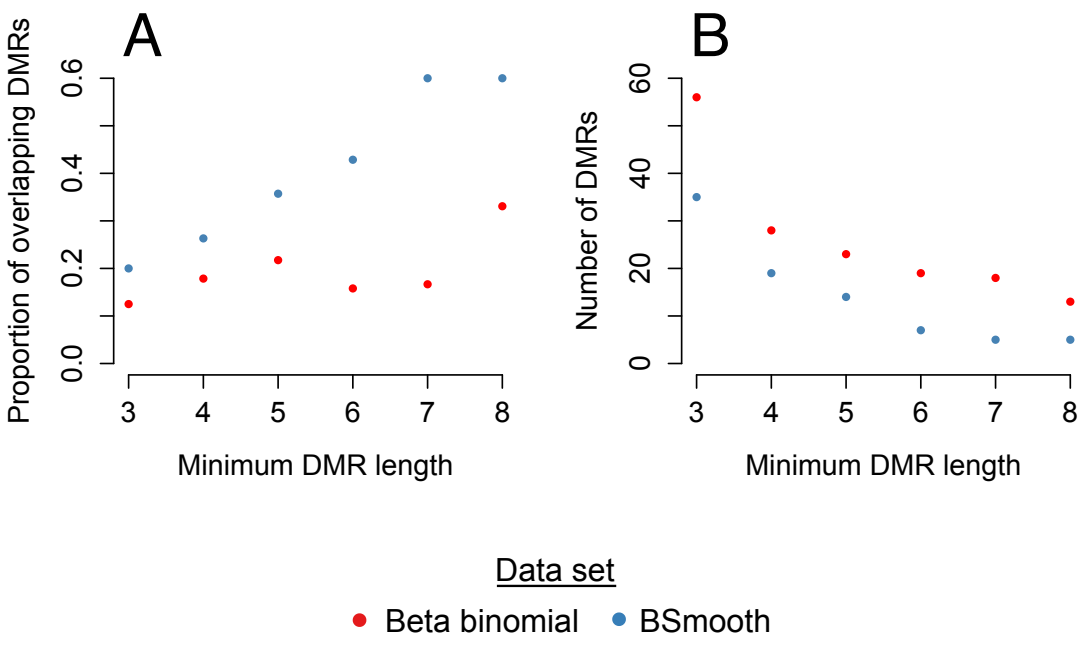


Supplementary Figure 8. Binarization of DNA methylation levels.
(A) Classification of simulated sites (with varying read depths and observed DNA methylation levels) as 'unmethylated' or 'methylated' based on a binomial test. Points are colored based on their classification at a given significance cutoff. For example, at α = 0.001, all sites with methylation level and read depth combinations represented by gray points or light blue points would be considered unmethylated, while all sites represented by medium or dark blue points would be considered methylated. (B) Technical properties of the data contribute to observations of 'sample-specific' methylation. For each combination of read depth and methylation level properties, we show the proportion of sites that were identified as exhibiting 'sample-specific methylation' (based on the definition used in[1]: ~0.6% of tested CpG sites). Results are shown for one representative focal sample versus the remaining seven samples, for read depth-methylation level combinations that occur at a minimum of five sites. Sites with high rates of sample-specific methylation were sequenced to higher coverage in the 'methylated' focal sample, compared to the mean coverage in the remaining seven 'unmethylated' samples (y-axis), producing an overall negative correlation. Additionally, the observed methylation level difference between the focal sample and the mean of the remaining seven samples is generally small on a continuous scale (x-axis).



Supplementary Figure 9. Agreement between "site-first" and "DMR-first" DMR identification approaches.

We simulated bisulfite sequencing count data using the correlation structure from a real RRBS data set[7] (across 1 chromosome), and called DMRs using *BSmooth* or an approach that aggregated across results from beta binomial models run on each site. (A) The proportion of called DMRs that overlapped with DMRs called using the alternative approach; results are thresholded based on the minimum number of CpG sites that occurred in the DMR in both data sets (x-axis). (B) The number of DMRs in each data set retained after thresholding by the minimum number of CpG sites. In this data set, the “site-first” method identifies a larger set of DMRs, and thus proportionally exhibits less overlap with the smaller set of *BSmooth* calls.



Supplementary References

1. Libbrecht, R., Oxley, P. R., Keller, L. & Kronauer, D. J. C. Robust DNA Methylation in the Clonal Raider Ant Brain. *Curr. Biol.* **26,** 1–5 (2016).
2. Janowitz Koch, I. *et al.* The concerted impact of domestication and transposon insertions on methylation patterns between dogs and grey wolves. *Mol. Ecol.* **25,** 1838–1855 (2016).
3. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10,** 232 (2009).
4. Lea, A., Tung, J. & Zhou, X. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genet.* **11,** e1005650 (2015).
5. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43,** 768–75 (2011).
6. Tobi, E. W. *et al.* DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nat. Commun.* **5,** 1–13 (2014).
7. Lea, A. J., Altmann, J., Alberts, S. C. & Tung, J. Resource base influences genome-wide DNA methylation levels in wild baboons (Papio cynocephalus). *Mol. Ecol.* **25,** 1681–1696 (2016).
8. Dubin, M. J. *et al.* DNA methylation variation in Arabidopsis has a genetic basis and appears to be involved in local adaptation. *eLife* **4,** e05255 (2015).
9. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44,** 821–4 (2012).
10. Liu, Y., Siegmund, K. D., Laird, P. W. & Berman, B. P. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* **13,** R61 (2012).
11. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history SUPP. *Nature* **499,** 471–475 (2013).
12. Horton, M. W. *et al.* Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat. Genet.* **44,** 212–216 (2012).
13. Danecek, P. *et al.* The variant call format and VCF tools. *Bioinformatics* **27,** 2156–8 (2011).
14. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57,** 289–300 (1995).
15. Lyko, F. *et al.* The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLoS Biol.* **8,** (2010).
16. Bonasio, R. *et al.* Genome-wide and caste-specific DNA methylomes of the ants camponotus floridanus and harpegnathos saltator. *Curr. Biol.* **22,** 1755–1764 (2012).
17. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
18. Delignette-Muller, M. L. & Dutang, C. fitdistrplus : An R Package for Fitting Distributions. *J. Stat. Softw.* **64,** 1–34 (2015).
19. Dabney, A. & Storey, J. qvalue: Q-value estimation for false discovery rate control. R package version 1.43.0. (2015).
20. Foret, S. *et al.* DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proc. Natl. Acad. Sci.* **109,** 4968–4973 (2012).
21. Team, R. D. C. R: A language and environment for statistical computing. (2012).
22. Hansen, K., Langmead, B. & Irizarry, R. BSmooth : from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13,** R83 (2012).
23. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100,** 9440–5 (2003).
24. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41,** 200–209 (2012).

25. Hansen, K. D. *et al.* BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13,** R83 (2012).
26. Hebestreit, K., Dugas, M. & Klein, H. U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* **29,** 1647–1653 (2013).
27. Yu, X. & Sun, S. HMM-DM: Identifying differentially methylated regions using a hidden Markov model. *Stat. Appl. Genet. Mol. Biol.* **15,** 69–81 (2016).
28. Feng, H., Conneely, K. N. & Wu, H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.* **42,** 1–11 (2014).
29. Dolzhenko, E. & Smith, A. D. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* **15,** 215 (2014).
30. Sun, D. *et al.* MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.* **15,** R38 (2014).
31. Jühling, F. *et al.* Metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* **26,** 256–262 (2016).
32. Akalin, A. *et al.* methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13,** R87 (2012).
33. Li, S. *et al.* An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics* **14 Suppl 5,** S10 (2013).
34. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42,** 764–770 (2014).
35. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–D669 (2014).