

UC Davis

UC Davis Previously Published Works

Title

Comethyl: a network-based methylome approach to investigate the multivariate nature of health and disease

Permalink

<https://escholarship.org/uc/item/137250sd>

Journal

Briefings in Bioinformatics, 23(2)

ISSN

1467-5463

Authors

Mordaunt, Charles E
Mouat, Julia S
Schmidt, Rebecca J
et al.

Publication Date




2022-03-10

DOI

10.1093/bib/bbab554

Peer reviewed

Comethyl: a network-based methylome approach to investigate the multivariate nature of health and disease

Charles E. Mordaunt , Julia S. Mouat, Rebecca J. Schmidt  and Janine M. LaSalle 

Corresponding author: Janine M. LaSalle, Department of Medical Microbiology and Immunology, Genome Center, Perinatal Origins of Disparities Center and MIND Institute, University of California, Davis, CA, USA. Tel.: +1-530-754-7598. Fax: 1-530-752-8692. E-mail: jmlasalle@ucdavis.edu

Abstract

Health outcomes are frequently shaped by difficult to dissect inter-relationships between biological, behavioral, social and environmental factors. DNA methylation patterns reflect such multivariate intersections, providing a rich source of novel biomarkers and insight into disease etiologies. Recent advances in whole-genome bisulfite sequencing enable investigation of DNA methylation over all genomic CpGs, but existing bioinformatic approaches lack accessible system-level tools. Here, we develop the R package Comethyl, for weighted gene correlation network analysis of user-defined genomic regions that generates modules of comethylated regions, which are then tested for correlations with multivariate sample traits. First, regions are defined by CpG genomic location or regulatory annotation and filtered based on CpG count, sequencing depth and variability. Next, correlation networks are used to find modules of interconnected nodes using methylation values within the selected regions. Each module containing multiple comethylated regions is reduced in complexity to a single eigennode value, which is then tested for correlations with experimental metadata. Comethyl has the ability to cover the noncoding regulatory regions of the genome with high relevance to interpretation of genome-wide association studies and integration with other types of epigenomic data. We demonstrate the utility of Comethyl on a dataset of male cord blood samples from newborns later diagnosed with autism spectrum disorder (ASD) versus typical development. Comethyl successfully identified an ASD-associated module containing regions mapped to genes enriched for brain glial functions. Comethyl is expected to be useful in uncovering the multivariate nature of health disparities for a variety of common disorders. Comethyl is available at github.com/cemordaunt/comethyl with complete documentation and example analyses.

Keywords: DNA methylation, whole-genome bisulfite sequencing, epigenetics, epigenome, weighted gene correlation network analysis, systems biology, autism spectrum disorder

Introduction

Despite the exceptional promise of genetics in understanding human health and disease, individual genes do not act in isolation from each other or from outside influences [1]. For instance, human susceptibility to noncommunicable diseases such as diabetes, cancer or cardiovascular disease is highly variable due to behavioral, social and economic factors in addition to underlying health disparities and genetics [2, 3]. Recent advances in genome sequencing technologies have greatly expanded our potential to understand the complex relationships that influence health trajectories and outcomes. However, existing statistical and analytical tools are frequently overly reductionist and specialized at determining the role of a single gene or environmental factor in disease by making precarious assumptions about the lack of gene × environmental interactions [4]. Unlike the relatively

static genome or the highly dynamic transcriptome, the DNA methylome is poised at the interface of genetics and a variety of environmental influences, reflecting past cellular variations in gene expression [5–8]. Through whole-genome bisulfite sequencing (WGBS), the DNA methylome is becoming a tractable ‘gold mine’ for discovering novel biomarkers and improved insights into multivariate mechanisms of disease [9]. What is currently lacking for WGBS and other DNA methylation technologies, however, is an improved analytical pipeline that intersects gene-level investigations with systems-level integration of multivariate data.

We developed the user-friendly R package Comethyl, available at github.com/cemordaunt/comethyl, for integration of WGBS data with additional metadata related to experimental conditions, cell type heterogeneity and patient demographics. The weighted gene correlation network analysis (WGCNA) R package [10] was designed

Charles E. Mordaunt, Ph.D., developed Comethyl while a postdoctoral fellow in the Department of Medical Microbiology and Immunology at UC Davis. He is currently a computational biologist at GSK.

Julia S. Mouat is a doctoral student in the Integrative Genetics and Genomics graduate group at UC Davis with interests in health disparities and intergenerational epigenetic risk factors for autism spectrum disorders.

Rebecca J. Schmidt, Ph.D., is an associate professor of Public Health Sciences at UC Davis, with expertise in the use of epigenetics in epidemiology and neurodevelopmental disorders.

Janine M. LaSalle, Ph.D., is a professor of Medical Microbiology and Immunology, Co-Director of the Perinatal Origins of Disparities Center, and Deputy Director of the Environmental Health Sciences Center at UC Davis, with expertise in epigenomics and neurodevelopmental disorders.

Received: July 13, 2021. **Revised:** November 15, 2021. **Accepted:** December 4, 2021

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

to perform network analysis of transcriptome data, so the gene is the unit of analysis, but transcribed genes make up only ~2% of the human genome. In addition, WGBS data require appropriate aggregation and filtering methods that differ from gene expression data. In Comethyl, we therefore developed an interface that extends WGCNA to allow user-determined genomic regions of clustered CpGs to be investigated for inter-connectedness. Regional clusters of CpGs are frequently correlated in methylation levels and are biologically more informative than individual CpGs because they are less susceptible to stochastic variation. Regions can also be defined based on other functional annotations, including gene bodies, promoters and enhancers. Once genomic regions are selected, the Comethyl package then adapts the existing WGCNA approach to identify modules of interconnected genomic regions based on correlated methylation levels. Once assigned to modules, these genomic regions can be mapped to genes, explored for enrichments to regulatory annotations and gene functions, and correlated with individual traits from experimental metadata to provide insights into complex conditions.

Autism spectrum disorder (ASD) is a complex, heterogeneous genetic condition with predicted gene \times environmental interactions in perinatal life [11, 12]. Using a high-risk prospective human cohort, we recently published a cord blood WGBS analyzed using a differentially methylated region (DMR) approach that identified novel ASD-associated loci [13]. In order to test the utility of Comethyl in identifying disease-relevant associations, we re-analyzed male cord blood WGBS data and identified a novel module that was negatively correlated with later ASD diagnosis but not with unrelated cell type, demographic or experimental factors. Here, we outline our analytical approach in Comethyl applied to this real-world dataset and discuss how it can be applied more broadly.

Methods and results

Comethyl pipeline, data requirements and CpG filtering

The Comethyl pipeline has three major subsections, shown as columns in the flow chart in Figure 1. The first section involves loading cytosine reports as well as defining CpGs and regions to be included in downstream analyses. Principal components of the included regions are identified and automatically adjusted for. Next, a comethylation network is constructed to call comethylation modules and characterize sample and module correlations. Lastly, modules are tested for significant correlations with relevant traits and experimental variables, mapped to genes, and assessed for functional enrichments.

We explored the use of Comethyl in a previously published WGBS dataset of 74 male newborn cord blood samples, including those that were diagnosed with

ASD or classified as typically developing by 36 months of age (TD $n=39$, ASD $n=35$). WGBS read alignment and quality control was performed using the CpG_Me workflow [14–17], which includes the generation of a Bismark CpG report. In Comethyl, individual sample Bismark CpG reports are read into a single BSseq object. The user can then select cutoff values to filter the CpGs. The cutoffs pertain to sequencing coverage and percent of samples, both of which should be balanced with the total number of CpGs included in downstream analysis (Supplementary Figure S1A and Supplementary Table S1). With low-pass WGBS (3–8 \times average CpG coverage), using strict or invariant coverage thresholds greatly reduces the number of genomic CpGs assayed, particularly as sample size increases. In this dataset, CpGs were filtered to include those with a minimum 2 \times coverage in 75% of the samples in order to balance total CpG number with sufficient sequencing depth and sample representation. Because Comethyl focuses on region-level methylation, relatively sparse CpG-level counts are aggregated together into regions for improved accuracy, and high coverage at individual CpGs is less critical.

Calling and filtering of genomic regions

Since DNA methylation of individual CpG nucleotides is imprecise but clusters of CpGs are regionally correlated, Comethyl summarizes CpG methylation at the region level prior to WGCNA. In the ASD cord blood dataset, regions are defined as CpG clusters containing at least three CpGs separated by at most 150 bp, though this can be adjusted by the user. Region features, such as the total number of regions, the total width of regions and total CpGs, are then visualized in comparison with minimum coverage and methylation standard deviation cutoffs to aid the user in determining appropriate thresholds for region filtering (Supplementary Figure S1B, Figure 2 and Supplementary Table S2). In the case of the ASD cord blood data, regions were selected for those with at least 10 reads in all samples and a standard deviation >0.05 , resulting in 251 717 regions for further analysis (Table S3). The overall goal of this step is to reduce the total number of regions assayed to those with sufficient coverage and variation in methylation between samples, which are thus the most likely to exhibit inter-connectivity and informative relationships with traits of interest.

In addition to the approach of calling genomic regions by CpG location, Comethyl also includes the option to define regions based on functional annotations, such as CpG islands, gene bodies, enhancers, or a custom annotation. This allows the user to constrain modules to include network nodes based on predefined regions. Since gene body methylation can correlate positively with expression [18, 19], gene bodies were selected as alternative regions to explore in the same cord blood

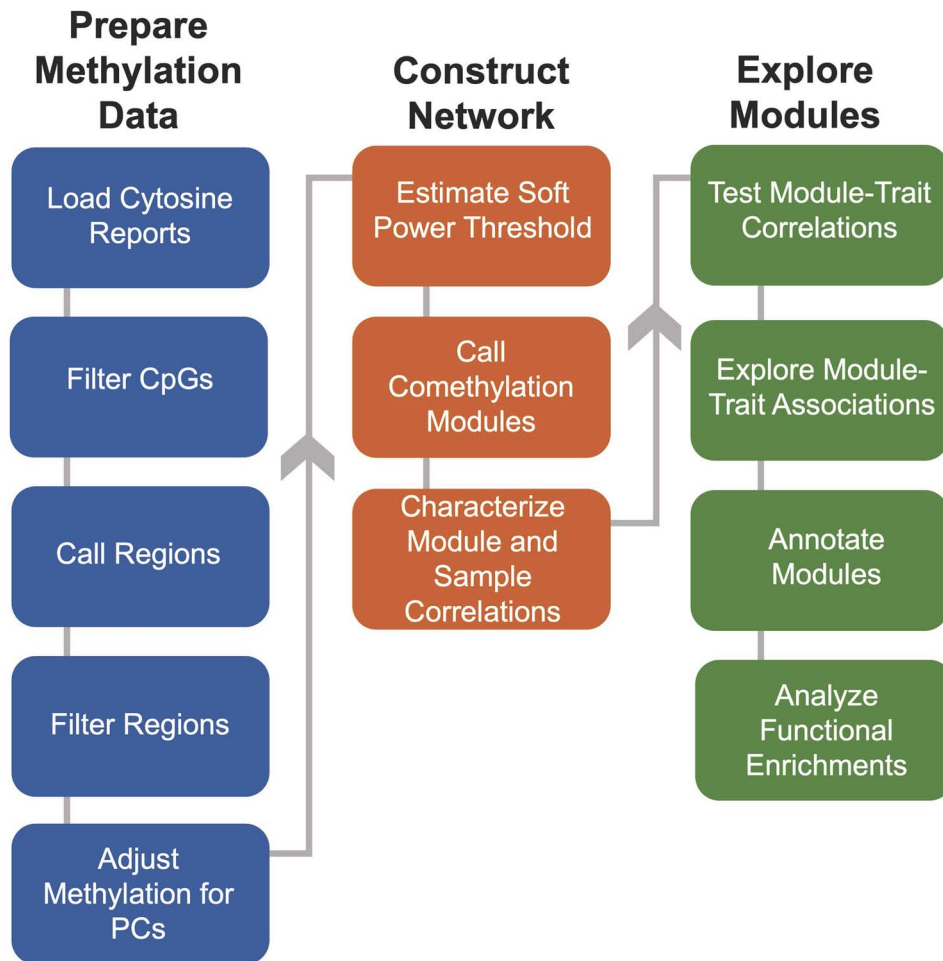


Figure 1. Comethyl pipeline for weighted region comethylation network analysis. Starting with CpG-level data from WGBS, the Comethyl pipeline consists of a suite of functions that allow users to summarize DNA methylation at the region-level in a biologically driven manner, construct a comethylation network, and explore associations of comethylation modules with traits.

dataset. Similar to the analysis of regions, gene bodies were then filtered for those with at least three CpGs and 10 reads in all samples (Supplementary Figure S2, Supplementary Tables S4 and S5).

Preparation for network construction

Before building the comethylation network, it is necessary to adjust the methylation data for confounding factors, check for outlier samples and evaluate scale-free topology. By default, percent methylation is directly calculated as the total methylated reads divided by the total reads in a region. One option is to instead use smoothed methylation data, for example using the BSmooth algorithm, which may be advantageous for low coverage datasets. Once the methylation values are obtained, Comethyl uses an unbiased principal component approach to adjust for the dominant variables in the data. This approach adapts `sva_network()` from the `sva` R package and is specifically designed to remove confounding technical variables and still allow for downstream network analysis. The required number of principal components is estimated and then the methylation data are adjusted. In the case of the cord

blood data, the top 10 principal components were used for adjustment. In order to check for outlier samples, the adjusted methylation data are used to cluster samples by Euclidean distance (Supplementary Figures S3A for regions, S4A for gene bodies). At this point, the input data for network analysis have been finalized and can be conceptualized as an $n \times m$ matrix $X = [x_{ij}]$, as defined in WGCNA [10]. Row indices are the network nodes or in this case regions (defined as $i = 1, \dots, n$), whereas column indices are sample measurements ($j = 1, \dots, m$).

Next, the scale-free topology is assessed with the goal of selecting a soft power threshold for network construction. The comethylation network is defined by its adjacency matrix a_{ij} , which contains the strength of the connection between each pair of nodes i and j . As implemented in WGCNA, we use a signed, weighted network with the adjacency matrix specified as $a_{ij} = (1 + \text{cor}(x_i, x_j)) / 2^\beta$ [10]. The soft power β is the power to which all correlations between genes are raised, a process that decreases background noise from weak correlations and amplifies stronger correlations. Soft power is plotted against the scale-free topology fit and mean

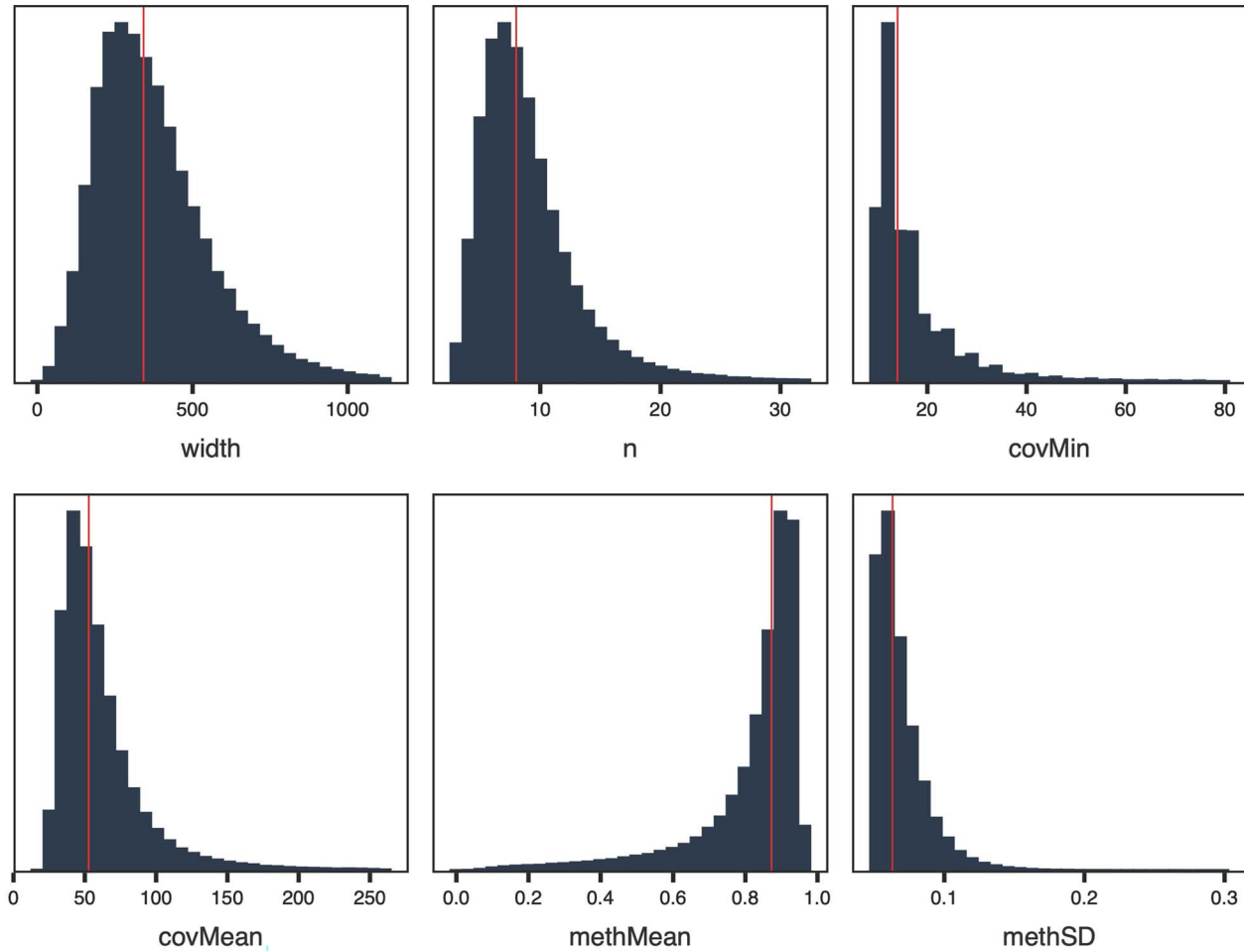


Figure 2. After filtering of CpG clusters, most regions have similar characteristics. CpGs with sufficient coverage in WGBS were grouped into clusters of at least three CpGs separated by at most 150 bp and filtered for those with at least 10 reads in all samples and standard deviation > 0.05 . Plots show the distributions of multiple region characteristics with the red line indicating the median value.

connectivity, which is the number of edges between nodes in the network and is inversely correlated with fit. It is recommended to select a soft power threshold that provides a fit of at least 0.8. The scale-free topology can be assessed with Pearson or biweight midcorrelation (Bicor) statistics. Pearson correlation is mean-based and more sensitive to outliers because it assumes that methylation data follow a normal distribution. In contrast, Bicor correlation is median-based and thus less sensitive to outliers; however, it also has reduced power to detect correlations. Because of the increased power with Pearson correlation, we used this for soft power estimation and network construction in the cord blood dataset. For datasets with high variability in methylation values, it may be more appropriate to use Bicor instead. It can be helpful to examine the modules and trait correlations observed for different network methods, in order to select the method that best fits the particular dataset. For the analysis of regions in the cord blood dataset, a soft-thresholding power cutoff of 18 was selected as the lowest power with a fit of at least 0.8 (Supplementary Figure S3B). For the analysis of gene body-defined regions, a soft power threshold of 12 was selected (Supplementary Figure S4B).

Network construction and calling of comethylation modules

The next step in Comethyl is to identify distinct groups of regions with correlated methylation values, defined as ‘comethylation modules’ using the network construction approach previously described for WGCNA [10]. Using the chosen soft power threshold and statistical model from the previous step, network construction uses a two-phase clustering approach to reduce computational intensity. In the first phase, regions are formed into large blocks using K-means clustering, where cluster centers are calculated with the first principal component and distances are based on correlation. In the second phase, an adjacency matrix, as described above, is calculated within each block and interconnectedness is assessed by the topological overlap measure [10]. Modules are detected based on hierarchical clustering followed by adaptive branch pruning. Modules from different blocks with highly correlated eigennodes can then be merged. Figure 3A shows the dendrogram with color-coded modules identified from one of eight blocks formed from the cord blood dataset. Regions not assigned to comethylation modules are assigned to the grey module. Interestingly, regions assigned

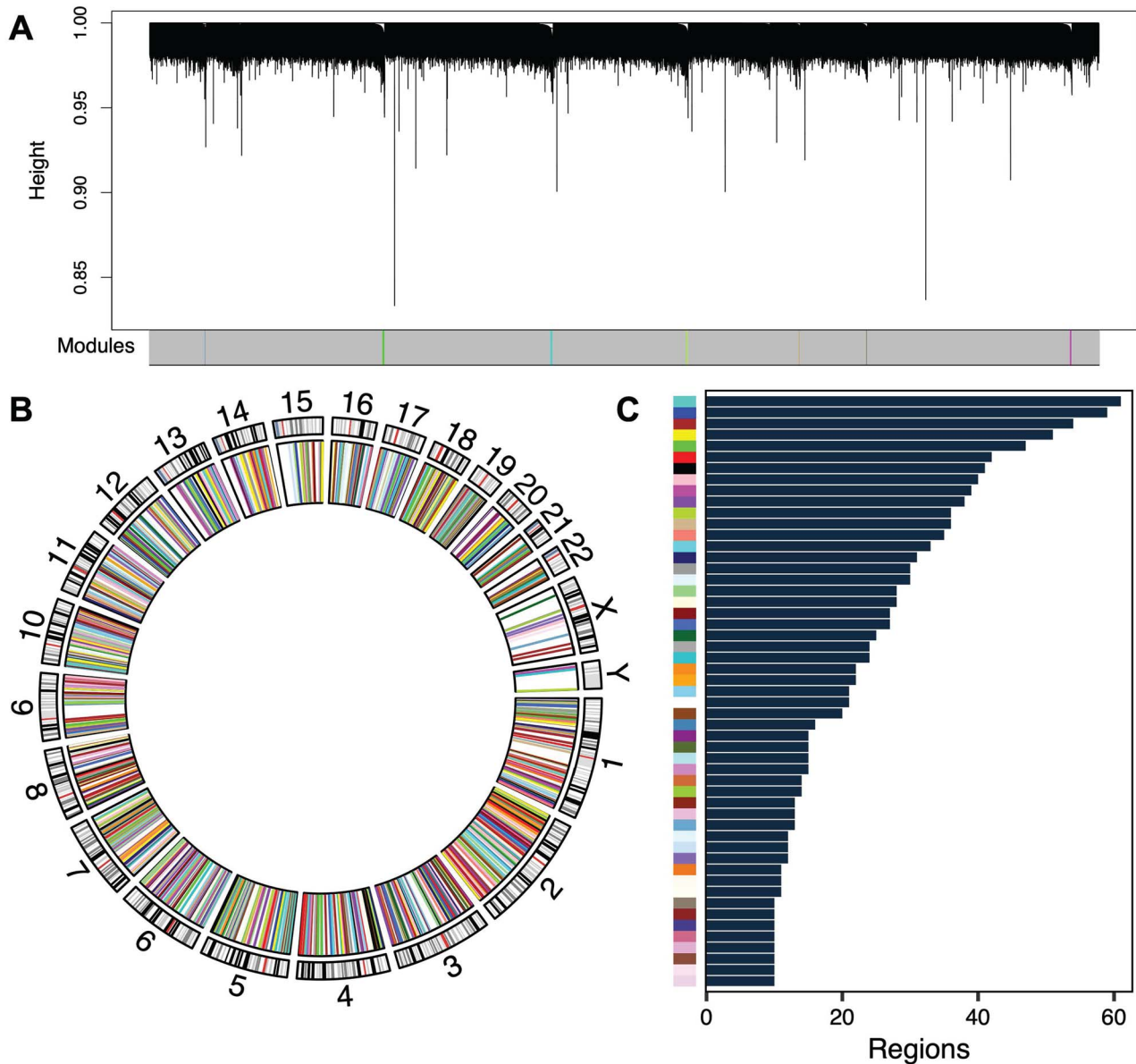


Figure 3. CpG cluster comethylation modules identified in cord blood. A comethylation network was constructed from adjusted region methylation data and assessed for modules of comethylated regions. (A) Region dendrograms for one block. (B) Circos plot of regions colored by module in the inner ring and chromosome bands in the outer ring. (C) Number of regions assigned to each module.

to a module make up a minority of cases. Regions within each module are distributed throughout the genome and may overlap multiple genes and/or CpG islands (Figure 3B). In this dataset, 53 modules were detected, and these ranged from the thistle2 module with 10 regions to the turquoise module with 61 regions (Figure 3C and Supplementary Table S3). With the gene body-based regions, 15 modules were identified, each including 10–139 regions (Supplementary Figure S5 and Supplementary Table S5).

Overall, we expect only a small number of regions (<1–2%) to be assigned to a module, at least for samples from a single tissue or cell type that come from a population with a complex disease. In contrast, a large portion of regions being assigned to a single module could indicate a confounding variable that has not been addressed

with principal component adjustments. Correlations of modules with samples and traits may reveal additional information to guide improved normalization.

Characterization of module and sample correlations

Eigenode values, which summarize the methylation values of each module, are calculated for each sample and then used to further examine relationships between samples and modules (Supplementary Figures S6 and S7). An eigenode represents the first principal component of the methylation matrix, defined as $E^{(q)}$ of the q th module [10]. An expected dynamic range of eigenode values was observed in the cord blood data, with samples clustering on the basis of higher or lower eigenode values of neighboring modules. Eigenode

correlations between samples and between modules are calculated with Pearson or Bicolor statistics, the latter of which was used for analysis of the cord blood dataset, and *P*-values are then derived based on Fisher's transformation. This approach can be used to assess interconnection between modules and identify highly related pairs (Supplementary Tables S6 and S7). Heatmaps of between-sample and between-module (Figures 4 and Supplementary Figure S8) correlations of eigennode values are also generated, revealing patterns of interconnectedness in the methylation data. Similarities based on module methylation between samples and modules can be investigated, potentially revealing subsets of related samples or modules.

Test all module-trait associations and explore informative modules

Once comethylation modules are identified and eigennode values defined for each module and sample, both continuous and discrete traits of the samples can be explored for module-trait associations. Sample traits have values for each column of matrix *X* and each trait is specified as a vector $T = (T_1, \dots, T_m)$ [10]. To calculate the eigennode significance, each sample trait is compared with each module eigennode by Pearson or Bicolor correlations, and *P*-values are calculated using Fisher's transformation. Sample traits can include all available information about potential variables of interest as well as potential confounding variables. In WGBS datasets, potential confounding variables include cell type proportions as well as technical variables including coverage, read duplication, read trimming and global cytosine methylation levels. For human populations, metadata should include clinical, diagnostic and demographic data, as well as sample collection characteristics, such as gestational age and birthweight for cord blood. For experimental studies in animal models or cell cultures, experimental variables should be included in the metadata for exploring module-trait relationships. One of the main features of this system-level approach is that correlations with many traits can be assessed simultaneously and then explored in greater detail.

Module-trait Bicolor correlations between all 49 variables and 53 modules from the unbiased region analysis in the cord blood dataset were tested (Figure 5A and Supplementary Table S8), and a number of module-trait associations were identified at a cutoff of $P < 0.05$ (Figure 5B). One comethylation module (Bisque 4) showed a highly significant negative correlation between sample eigennode values and ASD diagnosis ($P = 2.6E-4$). In other words, cord blood samples from newborns later diagnosed with ASD had lower Bisque 4 module eigennode values than those who were diagnosed as typically developing (Figure 5C). Other significant correlations with this module included behavioral and cognitive scores from the Autism Diagnostic Observation

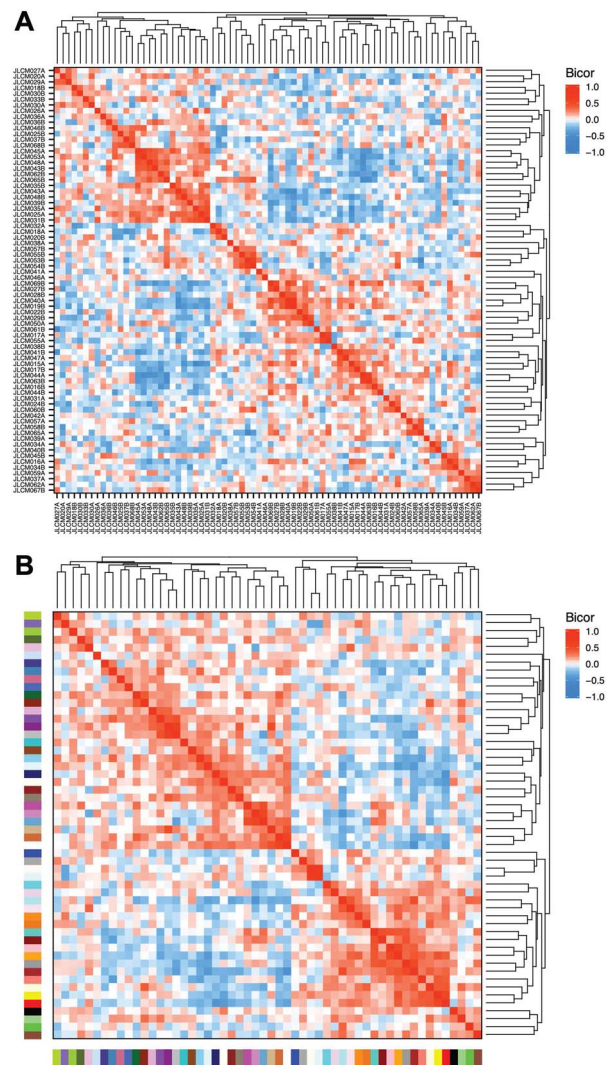


Figure 4. Correlations of CpG cluster module eigennode values reveal similarities between subsets of samples and modules. (A) Eigennode values were clustered and compared across samples using biweight midcorrelation. (B) Same as (A), but compared across modules.

Schedule (ADOS) and the Mullen Scales of Early Learning (MSEL), which were used in the diagnostic assessments for ASD. The Bisque 4 module was also correlated with gestational diabetes mellitus (GDM), which has been previously associated with increased risk for ASD [20, 21]. Most other comethylation modules were significant for associations with one or more traits, most frequently cell type proportions (granulocyte, B cell, nucleated red blood cell, CD4+ T cell) but also some technical variables (whole-genome % CpG or CpH methylation, or cohort study site). Assessing a range of variables is useful in determining which modules are of the greatest interest for further investigation, and which are confounded by other factors.

In the analysis of the gene body-defined regions of the same cord blood dataset, the black module was associated with ASD diagnosis and ADOS and MSEL scores; however, this module was more strongly correlated with

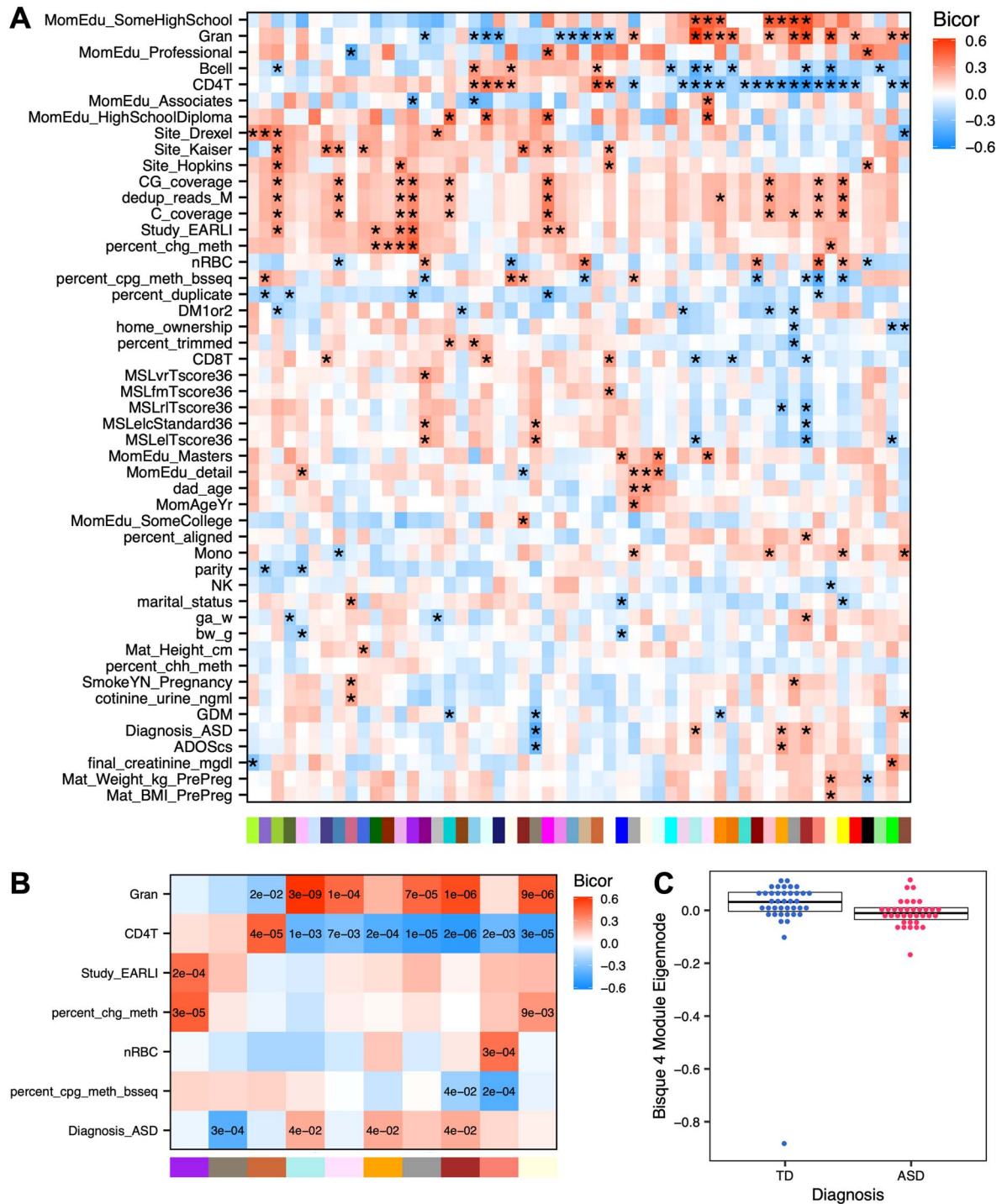


Figure 5. Correlations of CpG cluster module eigennodes with sample traits. (A) All tested correlations between module eigennode values and sample traits using biweight midcorrelation (* unadjusted $P < 0.05$). (B) Same as (A) but highlighting top significant module-trait associations. (C) Bisque 4 module is associated with later ASD diagnosis. Eigennode values are plotted for each sample in relation to ASD diagnosis. Box indicates 1st quartile, median, and 3rd quartile.

cell type proportions and whole-genome CpG methylation (Supplementary Figure S9 and Supplementary Table S9). Another example is the Green-Yellow module, which was associated with the discrete demographic variable of Home Ownership, as well as the continuous variable of cell type percentage for granulocytes, CD4+

T cells and CD8+ T cells (Supplementary Figure S10). More specifically, higher Green-Yellow module eigennode values were associated with home ownership, lower levels of granulocytes, and higher levels of CD4+ and CD8+ T cells. Weaker correlations were also found with maternal education, maternal height and one of the

Table 1. Regions and nearby genes identified in the Bisque 4 module associated with ASD diagnosis

chr	Start	End	CpGs	hubRegion	gene_symbol	distance_to_TSS	gene_context	CpG_context
chr1	57 479 875	57 479 935	7	TRUE	DAB1	-55 848	Intron	Open sea
chr2	224 079 272	224 079 787	9	FALSE	SERPINE2, FAM124B	-40 726, 322 504	Intergenic	Open sea
chr5	94 040 899	94 041 136	7	FALSE	POU5F2, FAM172A	-299 381, 70 617	Intron	Open sea
chr6	47 424 495	47 424 763	4	FALSE	TNFRSF21, CD2AP	-114 724, -53 160	Intergenic	Open sea
chr10	35 134 739	35 134 967	7	FALSE	CCNY, CREM	-202 021, 7393	Promoter, 1to5kb, intron	Open sea
chr11	80 720 879	80 721 186	14	FALSE	NA	NA	Intergenic	Open sea
chr13	52 925 862	52 926 068	11	FALSE	OLFM4, PCDH8	-102 794, -77 325	Intergenic	Open sea
chr15	68 240 587	68 240 932	12	FALSE	FEM1B, CLN6	-37 043, -11 017	Intron	Open sea
chr15	99 715 416	99 715 665	8	FALSE	LYSMD4, MEF2A	17 880, 149 454	Exon, intron, 3UTR	Open sea
chr19	43 838 028	43 838 499	6	FALSE	ZNF283, ZNF404	10 972, 45 700	Exon, intron, 3UTR	Open sea

Table 2. Gene bodies within the Green-Yellow Module associated with home ownership and three cell types

chr	Start	End	CpGs	hubRegion	strand	gene_symbol	gene_description
chr1	53 506 237	53 738 106	3113	FALSE	-	GLIS1	GLIS family zinc finger 1
chr3	6 770 001	7 741 533	4272	FALSE	+	GRM7	Glutamate metabotropic receptor 7
chr3	28 576 175	30 010 391	5242	FALSE	+	RBMS3	RNA binding motif single stranded interacting protein 3
chr4	92 303 966	93 810 157	4529	FALSE	+	GRID2	Glutamate ionotropic receptor delta type subunit 2
chr6	118 959 763	119 149 387	1240	FALSE	-	FAM184A	Family with sequence similarity 184 member A
chr7	101 815 904	102 283 958	8640	FALSE	+	CUX1	Cut like homeobox 1
chr11	43 358 920	43 494 933	681	FALSE	+	TTC17	Tetratricopeptide repeat domain 17
chr14	32 934 933	33 804 176	4769	FALSE	+	NPAS3	Neuronal PAS domain protein 3
chr14	89 124 871	89 954 659	8347	TRUE	-	FOXN3	Forkhead box N3
chr17	9 021 510	9 244 000	3669	FALSE	+	NTN1	Netrin 1
chr18	48 026 672	48 410 752	3843	FALSE	-	ZBTB7C	Zinc finger and BTB domain containing 7C
chr19	2 100 988	2 164 468	2088	FALSE	-	AP3D1	Adaptor related protein complex 3 subunit delta 1
chr19	31 149 979	31 349 436	2114	FALSE	-	TSHZ3	Teashirt zinc finger homeobox 3
chr20	57 648 392	57 711 536	1691	FALSE	-	PMEPA1	Prostate transmembrane protein androgen induced 1

MSEL scores. Notably, association with cell type proportion does not necessarily exclude a module from further consideration, but it does add important biological context for the interpretation of results.

Annotate modules and test for functional enrichments

The last step in Comethyl is to annotate genes to module regions and test enrichment of genes within each module for specific functions. For modules defined by the analysis of unbiased genomic regions, genes are mapped using Genomic Regions Enrichment of Annotations Tool (GREAT), gene information is added from BioMart and both gene and CpG island context is added from annotatr. The Bisque 4 module contains 10 regions, which map to 17 genes because most are in intergenic regions and all are in 'open sea' locations with respect to CpG islands (Table 1). For gene body-defined regions, genes are by definition already annotated. Table 2 shows an example

of the output table from the Green-Yellow module associated with home ownership, containing 14 gene bodies.

Functional enrichment analysis in GREAT is performed on regions within each module compared to all regions tested using the hypergeometric test and examined for overrepresentation of genes in Gene Ontology biological functions, cellular components and molecular functions, along with both human and mouse phenotype-associated genes. For the Bisque 4 module associated with ASD diagnosis, mapped genes were enriched in functions related to glial cells and glycosphingolipid metabolism, relevant to cerebellar development and thus ASD (Supplementary Figure S11 and Supplementary Table S10). For the Green-Yellow gene body module associated with home ownership and immune cell types, the 14 genes were enriched in glutamate receptor activity and several abnormal phenotypes, including abnormal respiration, postnatal lethality and increased prepulse inhibition (Supplementary Figure S12 and Supplementary Table S11). Functional enrichment

can be used to characterize the potential regulatory role of comethylation modules.

Assess module quality and preservation in an independent dataset

After module identification and functional association has been completed within a dataset, an important next step is the analysis of module quality and preservation in an independent replication dataset. One approach is that taken by Langfelder *et al.* [22], where multiple statistics are calculated to assess different types of preservation based on overlapping module assignments and network connectivity features. Observed statistics are then compared with an empirical null distribution, generated by permuting module assignments 100 times, in order to calculate Z-scores and P-values. In simulation studies by Langfelder *et al.*, Z-scores >2 were estimated as weak-to-moderate evidence for preservation, while Z-scores >10 were estimated as strong evidence.

Here, we compared our CpG cluster and gene body networks to those constructed using a second WGBS dataset in cord blood from males with ASD and controls (TD $n=17$, ASD $n=21$; [13]). Networks were built using the Comethyl pipeline and based on the same set of CpGs and regions, with some filtering for sequencing depth and variation. Preservation statistics were then calculated using the `modulePreservation()` function from WGCNA and visualized. As given by the `summary.qual` Z-score, all CpG cluster modules were scored as having strong evidence for good quality, except for the Ivory module, which had weak/moderate evidence (Supplementary Figure S13 and Supplementary Table S12). Similarly, all gene body modules were good quality (Supplementary Figure S14 and Supplementary Table S13). Using the cross-tabulation derived accuracy Z-score, 40/53 CpG cluster modules and 7/15 gene body modules had at least some evidence for preservation. In contrast, relatively few modules had evidence for preservation across these two datasets based on the network-derived `summary.pres` Z-score. Preserved modules with this statistic included the Light-Cyan, Blue and Floral-White CpG cluster modules and the Brown, Pink and Red gene body modules. This example demonstrates the value of multiple measurements of module quality and preservation in identifying robust and reproducible comethylation modules.

Discussion

Comethyl was designed as a user-friendly R package for analyzing WGBS data from human studies or relatively complex experimental animal models where multiple variables are potentially associated with or interconnected in influencing DNA methylation levels. An alternative approach to this problem is currently available in the Coordinate Covariation Analysis (COCOA) computational framework that uses covariation of traits

with epigenetic signals (including DNA methylation) across individuals and a predefined database of annotated region sets [23]. However, Comethyl is distinct from COCOA in the data-driven approach to defining region sets. Sets of comethylated regions are defined in the same dataset, rather than in datasets from completely different tissues or populations. The avoidance of genomic regions predefined from prior studies of human populations of predominantly European ancestry is an important goal in health disparities research [24, 25] for which Comethyl is better suited than COCOA. A second important distinction is that Comethyl uniquely uses network-based approaches to reduce the complexity of the data, which is critical for sequencing-based methylome data that includes millions of CpG sites. COCOA scores predefined region sets for how well they match variation in a phenotype. In contrast, Comethyl identifies precise modules of regions and then uses the resulting eigennode value of each module to perform correlations with multiple variables including the phenotypes of interest and technical factors. Because of this systems-based approach to simplify methylation data so that downstream multivariate analyses can more easily be applied, Comethyl is an improvement over COCOA for DNA methylation analyses.

Although WGCNA approaches have been utilized in the analysis of genome-wide DNA methylation previously [26–33], Comethyl appears to be the first one that utilizes the sequencing-based merits of WGBS data. In order to provide the most unbiased coverage of all genomic regions, including those over poorly annotated intergenic regions, Comethyl is specifically designed to empirically define and filter the regions that go into downstream module assignments. In addition to identifying regions agnostic to functional annotations, the Comethyl pipeline includes the flexibility for a user to conduct a focused analysis through the selection of gene bodies, CpG islands or other annotations as regions. A user could additionally integrate any type of region-based annotation with Comethyl, such as those from ChromHMM chromatin segmentations or ATAC-seq open chromatin peaks.

Comethyl was designed to work predominantly on low-pass WGBS data based on regional groups of CpGs similar to DMRs. If Comethyl is used to analyze EPIC/450k data, it is recommended to create regions of clustered CpGs or limit analysis to gene promoters or gene bodies. Alternatively, an approach developed for array data, such as CoMeBack, could be used [34]. The major limitation with array data is the sparse representation of CpGs over most of the genome and the difficulty in creating informative regions from biased CpG sampling [35, 36]. For instance, an array-based approach that selected CpGs based on predetermined gene body- and gene promoter-defined regions was used to identify modules associated with ASD diagnosis in a human brain study, but the gene functions identified were primarily associated

with immune, not brain function [26]. In contrast, our sequencing-based analysis of cord blood samples from newborns who were later diagnosed with ASD identified differentially methylated genes involved in brain development and glial functions [13] that have genetic overlap with ASD as well as epigenetic overlap with ASD brain and placenta [7, 37]. This discrepancy in results suggests that biased CpG coverage in array-based analyses may limit the ability to detect epigenetic changes in intergenic regions of the genome, important in developmental regulation of gene expression. The broad CpG coverage of WGBS-based Comethyl is especially important for identifying potential epigenetic biomarkers of disease for which the primary tissue affected, such as brain for ASD, is not able to be sampled in living children. In addition, the region-level approach in Comethyl aggregates low-pass WGBS CpG-level data into regions to improve accuracy without requiring excessive sequencing depth. Since the majority of WGBS-covered regions lie outside of those covered on EPIC and 450k arrays, they have been missed by array-based methylome studies. To avoid health disparities that can arise from the exclusive use of genomic platforms designed to human subjects of European ancestry [24, 25], sequencing-based approaches including WGBS and Comethyl offer the ability to make discoveries with broader relevance to human populations of diverse ancestry as well as to genomic regions that are poorly annotated.

Comethyl was designed to be complementary to DMR-based approaches to understand DNA methylation differences associated with diseases and exposures. Comethyl may be used in conjunction with DMR analysis to examine the correlated nature of CpG sites and their functional enrichments. The strength of DMR analyses is that they are more likely to identify specific individual genetic loci of high significance. However, the weakness is that some potential biologically relevant DMRs could be lost to noise in the underlying data and interacting variables affecting methylation levels. In contrast to DMR analysis, Comethyl does not consider absolute methylation values of the samples to group CpG sites. Instead, Comethyl allows the user to set windows or group adjacent CpGs, within which directional changes in methylation are used to group regions into modules. By correlating user-defined windows of the genome, Comethyl can analyze samples that have different absolute methylation values, which is prevented by other approaches such as DMR analysis. Through identifying sets of regions whose methylation varies together across samples, Comethyl may also identify regions with functionally related genes and link them to a sample trait. Another major advantage of Comethyl over DMR approaches is the ability to simultaneously examine all potential confounding associations with technical, experimental and demographic variables with the same comethylation modules as those with traits of interest. It is therefore recommended that Comethyl precede DMR analyses of WGBS data in complex datasets, so that

confounding variables can be identified and adjusted for in DMR calling. Traits can then be prioritized for those most likely to be associated with altered DNA methylation. Further, Comethyl has the potential to reveal important insights into health disparities in the data, such as the association between home ownership, considered a surrogate of socioeconomic status, and T cell and granulocyte proportions with the methylation at a set of genomic regions, which was observed in our analysis.

Integration of DNA methylation data with other large-scale multi-omics datasets is a major challenge for the future [38]. Comethyl reduces the complexity of the methylation analysis of 29 million CpG sites assayed to 20–80 distinct modules for which a single eigenvalue is assigned per sample. In this way, comethylation modules associated with the disease or trait of interest can be compared with additional measurements of phenotypes, metabolites and microbiota. Comethyl can be used in conjunction with gene expression analysis by comparing the gene body methylome with the transcriptome in the same sample population. This comparison allows for greater insight into epigenetic regulation of gene pathways and their role in disease, as well as the validity of epigenetic biomarkers in various biological contexts. Because Comethyl covers both coding regions and noncoding regulatory regions of the genome, it is highly relevant to the interpretation of common human polymorphisms identified from genome-wide association studies (GWAS). Genes and genomic regions annotated to comethylation modules that are significantly correlated with a trait of interest, such as diagnosis, can be compared with those identified by GWAS for potential overlap and examined for genetic effects on DNA methylation. Lastly, regions defined for Comethyl analysis can be defined from additional layers of epigenetic information, such as chromatin state maps (ChromHMM) defined from histone modifications, open chromatin and chromatin loop binding sites.

There are several potential limitations of Comethyl that should be considered during analysis. First, since most regions map distal to known genes, there is only an indirect link to the closest mapped genes, which may not be those functionally influenced by the methylation changes. Additional chromatin contact datasets or follow up experiments are needed to confirm gene mappings. Second, some biologically relevant methylation changes may be missed in regions filtered out based on low CpG density, low methylation variability or overly aggressive principal component adjustments. Therefore, a complementary approach of both Comethyl and DMR analyses is recommended to balance system-level with targeted analysis and find important associations that were missed with the alternate approach. In addition, replication in multiple datasets is required to identify reproducible modules and validate module-trait associations, especially since conventional multiple hypothesis testing adjustment methods are not compatible with WGCNA

[10]. Module reproducibility assessment is particularly important for datasets with small sample sizes and may be examined through multiple criteria, including cross-tabulation based and network based preservation statistics [22]. Cross-tabulation based approaches seek to identify corresponding modules between discovery and replication sets by overlapping module assignments. On the other hand, network-based approaches to assess preservation examine density (degree of connection between module nodes in discovery versus replication sets), connectivity (degree of similarity between node connection patterns in discovery versus replication sets), and separability (degree of distinction between modules in discovery versus replication sets). If the user is seeking to reproduce modules across datasets with similar properties (e.g. tissue, sex and species), care should be taken to reduce variability between replication cohorts in technical factors such as sequencing depth and platform along with biological differences between populations. Reproducibility measures may also be used to determine which modules are maintained across datasets of different tissues, sexes, or species, and which are specific to certain contexts [22].

Future goals for analyses with Comethyl include adding measured environmental exposures, polygenic risk scores, and circadian rhythmicity as traits to be correlated with modules. Including specific environmental exposures as traits would allow for direct correlation with epigenetic patterns regulating gene modules, as well as elucidating potential gene by environment interactions. Examining the functional enrichments of modules significantly correlated with exposures of interest could further suggest potential mechanisms of action. Adding polygenic risk score as a trait in a study of complex disease would allow for integrated analysis of genetic risk, environmental variables and epigenetic patterns. It would be enlightening, for example, if polygenic risk and a specific environmental variable were both significantly associated with a comethylation module. This could generate hypotheses about gene-environment interactions and disease etiology that could be explored in later functional studies.

Key Points

- We developed an R package, Comethyl, to bring multiple tools together and enable comethylation network analysis from low-pass WGBS data, which is available at github.com/cemordaunt/comethyl with complete documentation and example analyses.
- Comethyl is a complete workflow, from CpG-level counts to network construction, to functional enrichment of comethylation modules.
- We applied Comethyl to an ASD cord blood dataset with two different region methods and identified comethylation modules associated with traits of interest, including ASD, and enriched in relevant biological functions.

- Comethyl allows for system-level investigation into the impact of diverse traits on the methylome and on gene regulation.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Data availability

The data underlying this article are available in the Gene Expression Omnibus at <https://www.ncbi.nlm.nih.gov/geo/> and can be accessed with accession number GSE140730.

Acknowledgements

The authors thank Dr Benjamin Laufer for WGBS computational biology expertise, Dr Blythe Durbin-Johnson and Dr Kari Neier for statistical expertise and collaborators in the MARBLES cohort (Dr Irva Hertz Picciotto, Dr Sally Ozonoff and Dr Cheryl Walker) for human samples and data.

Funding

This work was supported by the National Institutes of Health (R01 ES029213, R01 ES025574, UG3/UH3 OD023365, P50 HD103526, P30ES023513 to J.M.L. and R.J.S. and T32 ES007059 to J.S.M.).

References

1. L. M. Hernandez, D. G. Blazer, Institute of Medicine (US) Committee on Assessing Interactions Among Social, *Genetics and Health*. National Academies Press (US), 2006. (21 January 2021. [Online], date last accessed). Available: <https://www.ncbi.nlm.nih.gov/books/NBK19932/>.
2. Hobbs A, Ramsay M. Epigenetics and the burden of noncommunicable disease: a paucity of research in Africa. *Epigenomics* 2015;**7**(4):627–39. <https://doi.org/10.2217/epi.15.17>.
3. Budreviciute A, Damiati S, Sabir DK, et al. Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors. *Front Public Health* 2020;**8**:574111. <https://doi.org/10.3389/fpubh.2020.574111>.
4. Williams EG, Auwerx J. The convergence of systems and reductionist approaches in complex trait analysis. *Cell* 2015;**162**(1):23–32. <https://doi.org/10.1016/j.cell.2015.06.024>.
5. Dunaway KW, Islam MS, Coulson RL, et al. Cumulative impact of polychlorinated biphenyl and large chromosomal duplications on DNA methylation, chromatin, and expression of autism candidate genes. *Cell Rep* 2016;**17**(11):3035–48. <https://doi.org/10.1016/j.celrep.2016.11.058>.
6. Mordaunt CE, Kieffer DA, Shibata NM, et al. Epigenomic signatures in liver and blood of Wilson disease patients include hypermethylation of liver-specific enhancers. *Epigenet Chromatin* 2019;**12**(1):10. <https://doi.org/10.1186/s13072-019-0255-z>.

7. Zhu Y, Mordaunt CE, Yasui DH, et al. Placental DNA methylation levels at CYP2E1 and IRS2 are associated with child outcome in a prospective autism study. *Hum Mol Genet* 2019;**28**(16):2659–74. <https://doi.org/10.1093/hmg/ddz084>.
8. LaSalle JM. A genomic point-of-view on environmental factors influencing the human brain methylome. *Epigenetics* 2011;**6**(7):862–9. <https://doi.org/10.4161/epi.6.7.16353>.
9. Dirks RAM, Stunnenberg HG, Marks H. Genome-wide epigenomic profiling for biomarker discovery. *Clin Epigenetics* 2016;**8**. <https://doi.org/10.1186/s13148-016-0284-4>.
10. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**(1):559. <https://doi.org/10.1186/1471-2105-9-559>.
11. Chaste P, Leboyer M. Autism risk factors: genes, environment, and gene-environment interactions. *Dialogues Clin Neurosci* 2012;**14**(3):281–92.
12. Wang C, Geng H, Liu W, et al. Prenatal, perinatal, and postnatal factors associated with autism: a meta-analysis. *Medicine (Baltimore)* 2017;**96**(18):e6696. <https://doi.org/10.1097/MD.00000000000006696>.
13. Mordaunt CE, Jianu JM, Laufer BI, et al. Cord blood DNA methylome in newborns later diagnosed with autism spectrum disorder reflects early dysregulation of neurodevelopmental and X-linked genes. *Genome Med* 2020;**12**. <https://doi.org/10.1186/s13073-020-00785-8>.
14. Laufer BI, Hwang H, Jianu JM, et al. Low-pass whole genome bisulfite sequencing of neonatal dried blood spots identifies a role for RUNX1 in Down syndrome DNA methylation profiles. *Hum Mol Genet* 2021;**29**(21):3465–76. <https://doi.org/10.1093/hmg/ddaa218>.
15. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011;**27**(11):1571–2. <https://doi.org/10.1093/bioinformatics/btr167>.
16. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**(1):Art. no. 1. <https://doi.org/10.14806/ej.17.1.200>.
17. Ewels P, Magnusson M, Lundin S, et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**(19):3047–8. <https://doi.org/10.1093/bioinformatics/btw354>.
18. Ball MP, Li JB, Gao Y, et al. Targeted and genome-scale methylomics reveals gene body signatures in human cell lines. *Nat Biotechnol* 2009;**27**(4):361–8. <https://doi.org/10.1038/nbt.1533>.
19. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**(7271):315–22. <https://doi.org/10.1038/nature08514>.
20. Xiang AH, Wang X, Martinez MP, et al. Association of maternal diabetes with autism in offspring. *JAMA* 2015;**313**(14):1425–34. <https://doi.org/10.1001/jama.2015.2707>.
21. Gardener H, Spiegelman D, Buka SL. Prenatal risk factors for autism: comprehensive meta-analysis. *Br J Psychiatry J Ment Sci* 2009;**195**(1):7–14. <https://doi.org/10.1192/bjp.bp.108.051672>.
22. Langfelder P, Luo R, Oldham MC, et al. Is my network module preserved and reproducible? *PLoS Comput Biol* 2011;**7**(1):e1001057. <https://doi.org/10.1371/journal.pcbi.1001057>.
23. Lawson JT, Smith JP, Bekiranov S, et al. COCOA: coordinate covariation analysis of epigenetic heterogeneity. *Genome Biol* 2020;**21**(1):240. <https://doi.org/10.1186/s13059-020-02139-4>.
24. Gurdasani D, Barroso I, Zeggini E, et al. Genomics of disease risk in globally diverse populations. *Nat Rev Genet* 2019;**20**(9):520–35. <https://doi.org/10.1038/s41576-019-0144-0>.
25. Popejoy AB, Ritter DI, Crooks K, et al. The clinical imperative for inclusivity: race, ethnicity, and ancestry (REA) in genomics. *Hum Mutat* 2018;**39**(11):1713–20. <https://doi.org/10.1002/humu.23644>.
26. Ramaswami G, Won H, Gandal MJ, et al. Integrative genomics identifies a convergent molecular subtype that links epigenomic with transcriptomic differences in autism. *Nat Commun* 2020;**11**(1):1. <https://doi.org/10.1038/s41467-020-18526-1>.
27. Sehl ME, Rickabaugh TM, Shih R, et al. The effects of anti-retroviral therapy on epigenetic age acceleration observed in HIV-1-infected adults. *Pathog Immun* 2020;**5**(1):291–311. <https://doi.org/10.20411/pai.v5i1.376>.
28. Busch R, Qiu W, Lasky-Su J, et al. Differential DNA methylation marks and gene comethylation of COPD in African-Americans with COPD exacerbations. *Respir Res* 2016;**17**. <https://doi.org/10.1186/s12931-016-0459-8>.
29. Massart R, Dymov S, Millecamps M, et al. Overlapping signatures of chronic pain in the DNA methylation landscape of prefrontal cortex and peripheral T cells. *Sci Rep* 2016;**6**. <https://doi.org/10.1038/srep19615>.
30. Tremblay BL, Guénard F, Lamarche B, et al. Network analysis of the potential role of DNA methylation in the relationship between plasma carotenoids and lipid profile. *Nutrients* 2019;**11**(6). <https://doi.org/10.3390/nu11061265>.
31. Dai Y, Lv Q, Qi T, et al. Identification of hub methylated-CpG sites and associated genes in oral squamous cell carcinoma. *Cancer Med* 2020;**9**(9):3174–87. <https://doi.org/10.1002/cam4.2969>.
32. Wong CCY, Smith RG, Hannon E, et al. Genome-wide DNA methylation profiling identifies convergent molecular signatures associated with idiopathic and syndromic autism in post-mortem human brain tissue. *Hum Mol Genet* 2019;**28**(13):2201–11. <https://doi.org/10.1093/hmg/ddz052>.
33. Chuang Y-H, Paul KC, Bronstein JM, et al. Parkinson's disease is associated with DNA methylation levels in human blood and saliva. *Genome Med* 2017;**9**. <https://doi.org/10.1186/s13073-017-0466-5>.
34. Gatev E, Gladish N, Mostafavi S, et al. CoMeBack: DNA methylation array data analysis for co-methylated regions. *Bioinformatics* 2020;**36**(9):2675–83. <https://doi.org/10.1093/bioinformatics/btaa049>.
35. Hurd PJ, Nelson CJ. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct Genomic Proteomic* 2009;**8**(3):174–83. <https://doi.org/10.1093/bfgp/elp013>.
36. Pidsley R, Zotenko E, Peters TJ, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* 2016;**17**(1):208. <https://doi.org/10.1186/s13059-016-1066-1>.
37. Vogel Ciernia A, Laufer BI, Hwang H, et al. Epigenomic convergence of neural-immune risk factors in neurodevelopmental disorder cortex. *Cereb Cortex* 2020;**30**(2):640–55. <https://doi.org/10.1093/cercor/bhz115>.
38. López de Maturana E, Alonso L, Alarcón P, et al. Challenges in the integration of omics and non-omics data. *Genes (Basel)* 2019;**10**(3):E238. <https://doi.org/10.3390/genes10030238>.