Gene set enrichment analysis for genome-wide DNA methylation data

Jovana Maksimovic[1,2,3], Alicia Oshlack[1,4], Belinda Phipson[1,2*]

[1] Peter MacCallum Cancer Centre, Melbourne, Victoria, 3000, Australia

[2] Department of Pediatrics, University of Melbourne, Parkville, Victoria, 3010, Australia

[3] Murdoch Children's Research Institute, Parkville, Victoria, 3052, Australia

[4] School of Biosciences, University of Melbourne, Parkville, Victoria, 3010, Australia

[*] corresponding author

Email addresses:

Jovana Maksimovic: Jovana.Maksimovic@petermac.org

Alicia Oshlack: Alicia.Oshlack@petermac.org

Belinda Phipson: Belinda.Phipson@petermac.org

Abstract

DNA methylation is one of the most commonly studied epigenetic marks, due to its role in disease and development. Illumina methylation arrays have been extensively used to measure methylation across the human genome. Methylation array analysis has primarily focused on preprocessing, normalisation and identification of differentially methylated CpGs and regions. GOmeth and GOregion are new methods for performing unbiased gene set testing following differential methylation analysis. Benchmarking analyses demonstrate GOmeth outperforms other approaches and GOregion is the first method for gene set testing of differentially methylated regions. Both methods are publicly available in the *missMethyl* Bioconductor R package.

Keywords

DNA methylation, gene set analysis, differential methylation, statistical analysis

Background

DNA methylation is essential to human development, with roughly 3-6% of all cytosines methylated in normal human DNA (Esteller 2007). Epigenetic marks can be modified by environmental exposures, and methylation changes are known to accumulate with age. Aberrant methylation patterning is associated with many diseases, which has led to several large studies profiling DNA methylation, such as The Cancer Genome Atlas, Encyclopedia of DNA Elements and numerous epigenome-wide association studies.

Both array and sequencing based technologies are available for profiling DNA methylation at a genome-wide scale. However, even though the cost of sequencing has dramatically decreased, the ease and cost effectiveness of the Illumina human methylation arrays have ensured that the array platforms remain a popular choice for many researchers. To date, a major focus when analysing DNA methylation data has been the identification of significantly differentially methylated CpG sites between groups of samples in a designed experiment. There are many well-established analysis methods that perform normalisation and statistical testing for this purpose, including publicly available software packages, such as *limma* (Ritchie et al. 2015), *minfi* (Aryee et al. 2014), *missMethyl* (Phipson, Maksimovic, and Oshlack 2016), *methylumi* (Davis et al. 2019), *wateRmelon* (Pidsley et al. 2013), *ChAMP* (Morris et al. 2014), *RnBeads* (Assenov et al. 2014; Müller et al. 2019), *Harman* (Oytam et al. 2016), and *ENmix* (Z. Xu et al. 2016) . It is well established that methylation of CpG sites is spatially correlated along the genome (Eckhardt et al. 2006) and that long tracks of differential methylation are often more biologically meaningful than differences at individual CpG sites (Hansen et al. 2011). This has led to region-based analyses, with Bioconductor R packages such as *Probe lasso* (Butcher and Beck 2015), *bumphunter* (Jaffe et al. 2012), *DMRcate* (Peters et al. 2015), *mCSEA* (Martorell-Marugán, González-Rumayor, and Carmona-Sáez 2019) and *DMRforPairs* (Rijlaarsdam et al. 2014) developed specifically for this purpose.

Once differential methylation analysis between groups of samples has been performed, there may be a long list of significant CpG sites or regions for the researcher to interpret. A popular approach to gain a more systems-level understanding of the changes in methylation is to examine which gene pathways may be enriched for differential methylation in the

experiment. This approach was popularised in the analysis of gene expression microarrays and RNA-sequencing (RNA-seq) data, with one of the first methods, GSEA, published in 2005 (Subramanian et al. 2005). Since then, a number of gene set testing methods have been developed (e.g. Yaari et al. (2013), Wu et al. (2010), and Wu and Smyth et al. (2012)), all of which are specific to gene expression data, with the GOSeq method (Young et al. 2010) developed specifically to account for gene length bias in RNA-Seq data.

Methylation, however, is a DNA mark that can occur anywhere on the genome and is not as directly related to genes as expression data. Therefore, a methylation specific issue in performing gene set testing is how to assign differentially methylated features to genes. Thus far, there are very few gene set testing methods designed specifically for DNA methylation data, and often ad hoc approaches are taken. Only two other methods, ebGSEA, available in the *ChAMP* R Bioconductor package (Dong et al. 2019), and *methylGSA* (Ren and Kuan 2018), have been specifically proposed for gene set enrichment analysis (GSEA) of methylation array data. *MethylGSA* is an R Bioconductor package that contains several different gene set testing approaches: mRRA, which adjusts multiple p-values for each gene by Robust Rank Aggregation followed by either over-representation analysis (ORA) or functional class scoring in combination with GSEA, and mGLM, which is an extension of GOglm, implementing a logistic regression to adjust for the number of probes in the enrichment analysis (Mi et al. 2012). The ebGSEA method uses a global test to rank genes, instead of CpGs, based on their total level of differential methylation; enrichment of gene sets is then calculated from the ranked gene list using either a Wilcoxon Test (WT) or Known Population Median Test (KPMT) (Dong et al. 2019). Both ebGSEA and the *methylGSA* methods use individual CpG probe-based differential methylation features, and we are presently not aware of any methods for performing gene set testing for differentially methylated regions.

Here we present GOmeth and GOregion to perform gene set analysis in the context of DNA methylation array data for differential methylation of CpG sites and regions, respectively. The key aspect of our methods is the ability to take into account biases inherent in the data, which relate to how differentially methylated probes are annotated to differentially methylated genes, that can then be assigned to a gene set. Specifically, measured CpG sites are not distributed evenly across the genome, and we and others show that genes that have

more CpG sites measured across them are more likely to be detected as differentially methylated compared to genes that have fewer measured CpG sites. In addition, approximately 10% of gene-annotated CpGs are assigned to more than one gene, violating assumptions of independently measured genes. We consider both of these biases in our methods for detecting enriched pathways based on probe-wise and region-based differential methylation analysis.

In this paper we have evaluated the performance of our methods on real and simulated data, as well as comparing to the other available methods across a variety of datasets. We found that our methods were the best statistical and computational performers across a variety of comparisons and gene set testing collections. For experiments with many thousands of significantly differentially methylated probes, we recommend GOregion for extracting the most biologically meaningful results from the data. Our methods are publicly available in the Bioconductor R package, *missMethyl.* All of the analysis performed in this paper can be found at the following website: http://oshlacklab.com/methyl-geneset-testing/. The GitHub repository associated with the analysis website is at: https://github.com/Oshlack/methyl-geneset-testing.

Results

Composition biases of 450K and EPIC arrays

Consider the scenario where we have performed differential methylation analysis on individual CpG sites. In order to perform gene set enrichment analysis based on the results from a probe-wise differential methylation analysis, we need to annotate each probe on the array to a gene. One approach for gene set testing is to simply call a gene differentially methylated if at least one CpG site associated with that gene is significantly differentially methylated, and this has been used in many previous analyses (e.g. Zhang et al. (2013), Phipson and Oshlack (2014)). The problem with this approach is that the numbers of CpG sites measured across each gene varies significantly across the genome (Figure 1A, Supplementary Figure 1A). For the 450K array, the minimum number of CpGs measured per gene is 1 and the maximum is 1299 with a median of 15, based on the

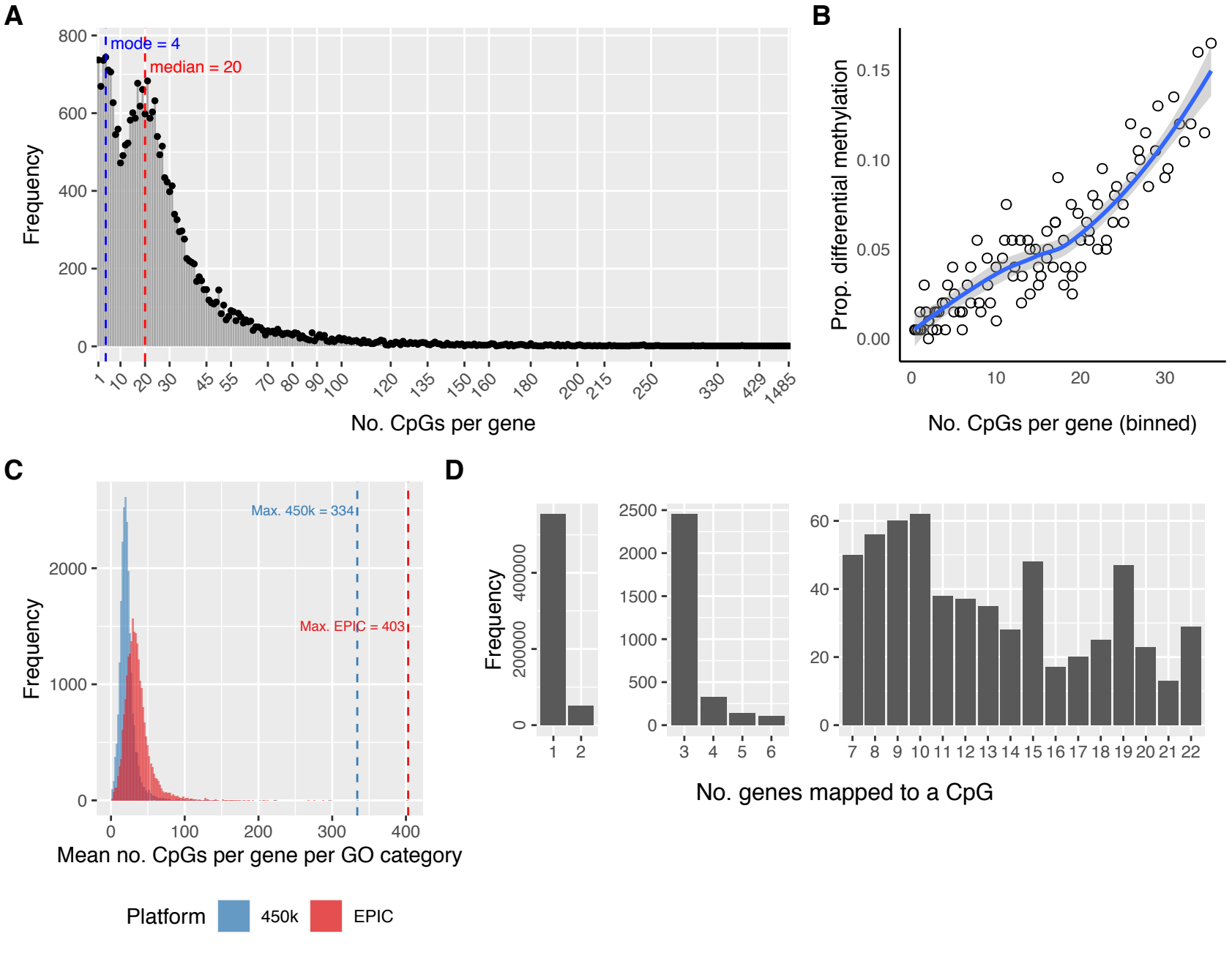

A
mode = 4
median = 20
Frequency
No. CpGs per gene
B
Prop. differential methylation
No. CpGs per gene (binned)
C
Max. 450k = 334
Max. EPIC = 403
Frequency
Mean no. CpGs per gene per GO category
Platform
450k
EPIC
D
Frequency
No. genes mapped to a CpG

Figure 1. **Array design bias for the Illumina HumanMethylation EPIC BeadChip. (A)** Frequency plot of the numbers of CpGs measuring methylation across each gene for the EPIC array. The most extreme value is 1485 CpGs measuring methylation across a single gene. The median is 20 and the mode is 4. **(B)** Plot demonstrating probe-number bias for B-cells vs. NK cells from sorted blood cell type EPIC array data. Genes with more measured CpGs are more likely to be differentially methylated. **(C)** Histogram of the median numbers of CpGs per gene for each GO category for the 450K and EPIC arrays. The distributions differ between the arrays, however, both show a varying number of CpGs per gene per GO category. GO categories with more CpGs per gene, on average, have greater power to be significantly enriched. **(D)** Split bar chart showing the numbers of genes annotated to each CpG (multi-gene bias). While the majority of CpGs are annotated to only one gene, there is still a large number annotated to 2 or more genes.

IlluminaHumanMethylation450kanno.ilmn12.hg19 annotation package. For the EPIC array, the numbers of CpGs measured across genes ranges from 1 to 1485 (median = 20, *IlluminaHumanMethylationEPICanno.ilm10b4.hg19* annotation package). Genes that have larger numbers of CpGs measured are more likely to be called differentially methylated when comparing B-cells vs natural killer cells (Figure 1B), and this holds true for the majority of data sets. This bias towards genes with more measured CpG sites can in turn influence the probability of a gene set being called significantly enriched, as some gene sets contain genes with more than the average number of CpGs, and some have genes with fewer measured CpGs (Figure 1C). In this paper, we refer to this particular source of bias as "probe-number bias".

Approximately 70% of the probes on the EPIC array are annotated to at least one gene (74% for 450K array). However, another annotation issue, which is more subtle, is that a single CpG may be annotated to more than one gene as the gene regions overlap on the genome. While the majority of CpGs with gene annotations are associated with only one gene (329,365/359,832 = 92% for 450K, 554,221/607,820 = 91% for EPIC arrays), there are still a large number of CpGs annotated to 2 or more genes (Figure 1D, Supplementary Figure 1B). This can cause issues as the measurements of differentially methylated genes are not independent. If we use every gene associated with a single CpG, we risk counting a single significant CpG site multiple times when including the genes as enriched in a gene set of interest. If these genes were evenly distributed across the GO categories, this may not be an issue. However, genes that are close in genomic proximity can be functionally related and be present in a single GO category. An extreme example of this is cg17108383 which is annotated to 22 genes, all belonging to the protocadherin gamma gene cluster (Supplementary Figure 2). All 22 of these genes are present in the GO category "GO:0007156: homophilic cell adhesion via plasma membrane adhesion molecules", which contains a total of 129 genes. If each of these significant genes are included when performing a hypergeometric test for enrichment of the GO category, then for this single significant CpG site, the overlap between the differentially methylated genes and the genes in the gene set is increased by 22, and this GO category will appear significantly enriched. Unfortunately, this is not an isolated occurrence. For the EPIC array, gene ontology (GO) enrichment analysis on the 53,599 CpGs that are annotated to at least 2 genes results in 114 significantly enriched

GO categories (Holm's adjusted p-value cut-off < 0.05). Restricting to CpGs that are annotated to at least 3 genes results in significant enrichment of 56 GO categories (Holm's adjusted p-value cut-off < 0.05), which are mostly related to processes involved in transcriptional regulation (Supplementary Table 1).

We refer to this newly identified source of bias as "multi-gene bias". In order to reduce false positives, it is important to take this multi-gene bias into account when calculating the intersection between differentially methylated genes and the genes in each gene set. One approach for dealing with multi-gene bias is to simply randomly select one gene to be represented by the CpG, but this approach risks losing valuable information by ignoring the remaining associated genes. We include this multi-gene bias in our statistical framework for gene set testing and ensure significant CpGs are only counted once, at most.

GOmeth performs gene set testing on differentially methylated CpG sites

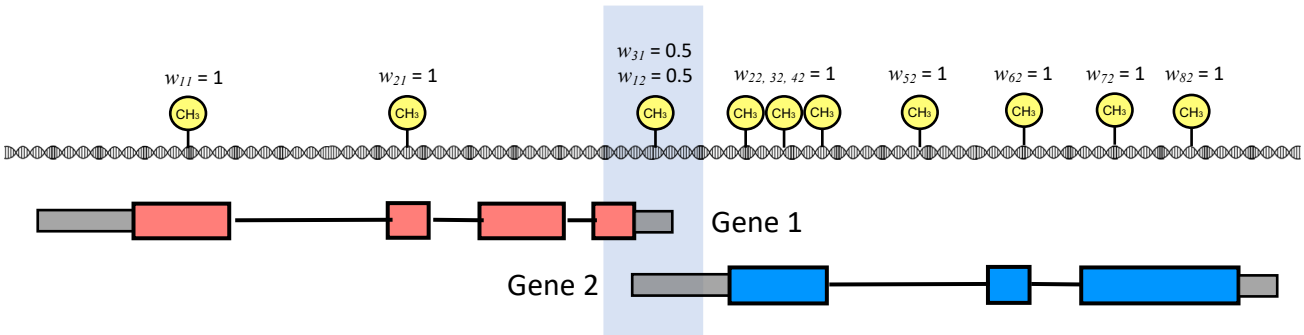
Our method for gene set testing performs enrichment analysis of gene sets while correcting for both probe-number and multi-gene bias in methylation array data. This method was inspired by the GOSeq method (Young et al. 2010). The GOSeq method was designed to account for the fact that longer genes have more sequencing read counts compared to shorter genes, and hence have more power to be statistically significantly differentially expressed. Similarly, we see that genes with a larger number of measured CpGs have a higher probability of differential methylation (Figure 1B). To account for this bias, GOmeth assigns a probability of differential methylation to each gene given the number of statistically differentially methylated CpGs in the data and the total number of annotated CpGs per gene. As the probabilities are calculated empirically from the data, this trend may look different between different datasets, but generally we always observe a strong positive trend. From this empirical trend, the odds of differential methylation for each gene set or gene ontology category is calculated. Specifically, a test based on Wallenius' noncentral hypergeometric distribution is performed for each gene set or gene ontology category, incorporating the odds that a gene set is more or less likely to be enriched based on the probe-number bias of the array (Phipson, Maksimovic, and Oshlack 2016).

As previously noted, another issue that needs to be addressed is the multi-gene bias where a CpGs is annotated to more than one gene.The solution we propose is to perform fractional counting by assigning a weight to each CpG that is dependent on how many genes are annotated to the CpG (Figure 2). For example, if a significantly differentially methylated CpG is annotated to 2 genes, each CpG is assigned a weight of 0.5 instead of one. If no other significant CpGs are annotated to that particular gene, then the gene will contribute a "count" of 0.5 to the intersection statistic for the Wallenius' noncentral hypergeometric test. Thus, if both genes are present in the same gene ontology category, they will contribute a total count of at most one to the intersection statistic. For genes with multiple significant CpGs that may include several multi-gene associated CpGs, the total count a gene can contribute to the intersection statistic is 1. By incorporating the fractional contribution from any multi-gene-associated CpG, we can also calculate the "equivalent" numbers of CpGs associated with each gene and incorporate this in the odds calculation for each gene set (Figure 2).

We have implemented our approach to gene set testing with two different functions in the *missMethyl* Bioconductor R package, 'gometh' and 'gsameth'. The difference between the two functions is minimal, with 'gometh' specifically testing for enrichment of gene ontology (GO) categories from the *GO.db* annotation package, or KEGG pathways from the *KEGG.db* annotation package. The 'gsameth' function is a more general version of 'gometh', where the user can supply any list of gene sets to be tested. In addition, the 'gometh' and 'gsameth' functions allow the set of significantly differentially methylated CpGs to be restricted to genomic regions of interest such as promoters or gene bodies, as these may interrogate different biological pathways.

Improved Type I error rate control with GOmeth

We first tested the performance of GOmeth by randomly sampling sets of CpG probes from the EPIC and 450K array annotation that we designated as differentially methylated and running gene ontology analyses. Under these null scenarios we would not expect to see significant enrichment of any GO categories. We randomly selected 100 sets each of 50, 100, 500, 1000, 5000 and 10,000 CpGs as "significantly" differentially methylated based on the



	Num CpGs	Equiv CpGs (N_j)
Gene 1	3	2.5
Gene 2	8	7.5

w_{ij} = weight for CpG i in gene j

Figure 2. Overview of how probe number and multi-gene bias is taken into account in GOmeth. CpGs are not evenly spaced throughout the genome. Gene 1 has methylation measured at three CpGs and gene 2 has methylation measured at eight CpGs. The CpG shaded in blue is an example of a shared genomic location between gene 1 and 2, with one CpG measuring the methylation status for two genes and is thus not independently measured. By calculating a weight for each CpG inversely proportional to how many genes that CpG is annotated to, we can calculate the equivalent numbers of CpGs measured across each gene and ensure the enrichment statistic for Wallenius' noncentral hypergeometric test is not artificially inflated due to multi-gene annotated CpGs.

450K and EPIC annotation. We tested for enrichment of gene ontology sets for each simulation and calculated the number of GO categories that were significant at a p-value threshold of 0.05. For a test to correctly control the type I error rate, we expect 5% or fewer GO categories to have significant p-values for random data. We compared the three testing options available in GOmeth: the hypergeometric test with no bias corrections ("HGT"), Wallenius' noncentral hypergeometric test taking into account probe-number bias only ("HGT-mod") and GOmeth, which takes into account both probe-number and multi-gene bias (Figure 3A & B, Supplementary Figure 3). Under these simulation conditions we were unable to compare to the *methylGSA* methods and ebGSEA since these tests require M-values or β values as input rather than just the list of significant probes.

Figure 3A summarises the type I error rates for our methods across the varying sets of randomly selected CpGs sampled from the EPIC array. As the numbers of differentially methylated CpGs increased we noted that the hypergeometric test reported too many significant GO categories, particularly for more than 500 CpGs. Clearly, taking into account the probe-number bias makes the biggest correction to false discoveries, with HGT-mod and GOmeth maintaining the correct Type I error rate. Correcting for multi-gene bias (GOmeth) further reduced the numbers of significantly enriched GO categories. We did notice that not accounting for the multi-gene bias led to an increase in false discoveries as the numbers of significant CpGs increased. A similar trend was observed for simulations based on the 450K array (Supplementary Figure 3A).

Of the significant GO categories reported by each method, we found that the hypergeometric test was more biased towards GO categories with more CpGs measured per gene on average (Figure 3B, Supplementary Figure 3B). HGT-mod and GOmeth reported significant categories that had a wide range of CpGs per gene on average. Based on the results of these simulations, we selected GOmeth taking into account both probe-number and multi-gene bias as the best option to use for further analysis, even though it is quite a conservative test in this particular simulation scenario.

The prior simulations are limited in that they only generate a random set of CpGs to test for over-representation of gene sets. However, previously published methods, ebGSEA and the *methylGSA* methods, require differential methylation measurements (M-values or β values)

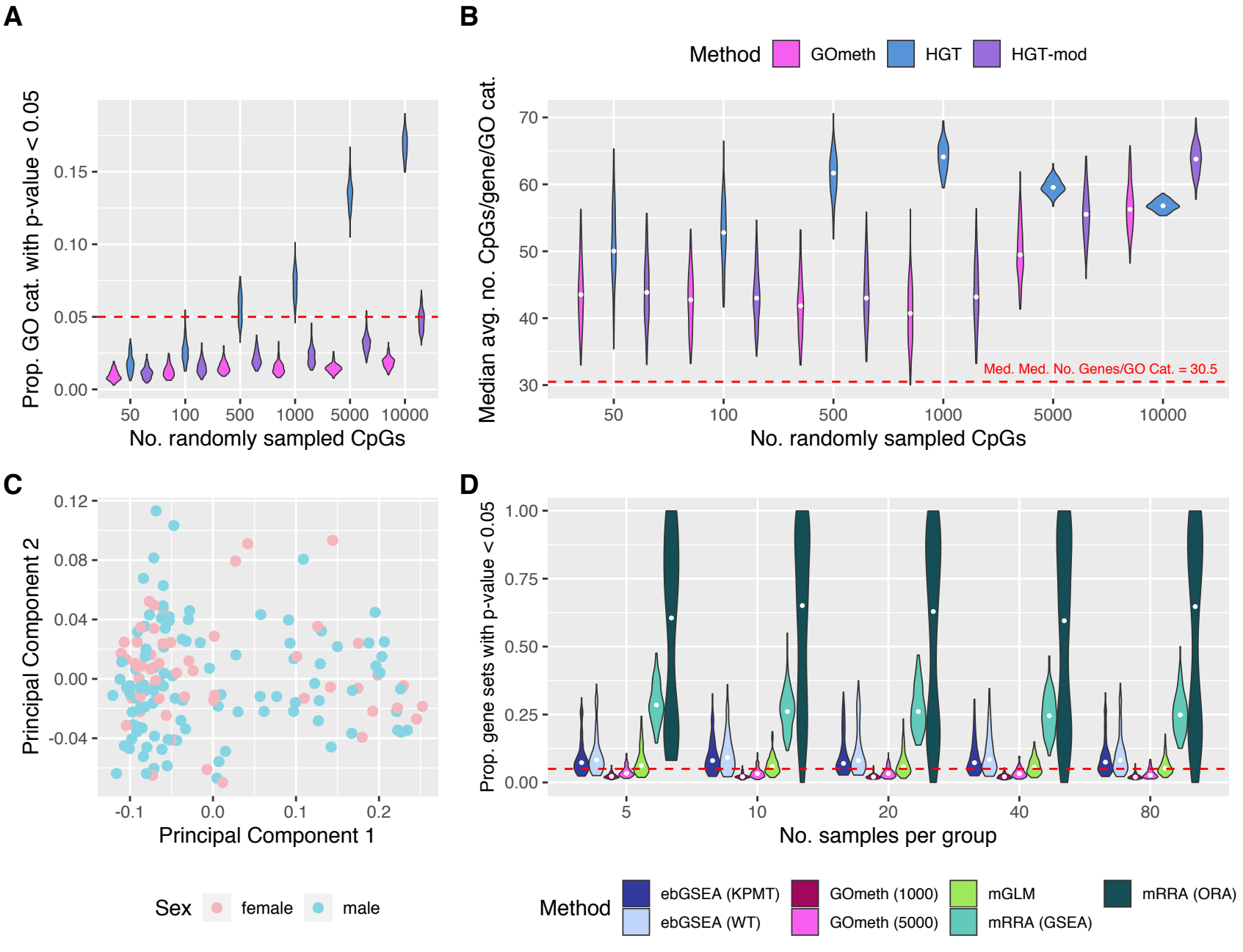

A
Prop. GO cat. with p-value < 0.05
0.15
0.10
0.05
0.00
50
100
500
1000
5000
10000
No. randomly sampled CpGs
B
Method
GOmeth
HGT
HGT-mod
Median avg. no. CpGs/gene/GO cat.
70
60
50
40
30
Med. Med. No. Genes/GO Cat. = 30.5
50
100
500
1000
5000
10000
No. randomly sampled CpGs
C
Principal Component 2
0.12
0.08
0.04
0.00
-0.04
-0.1
0.0
0.1
0.2
Principal Component 1
Sex
female
male
D
Prop. gene sets with p-value < 0.05
1.00
0.75
0.50
0.25
0.00
5
10
20
40
80
No. samples per group
Method
ebGSEA (KPMT)
ebGSEA (WT)
GOmeth (1000)
GOmeth (5000)
mGLM
mRRA (GSEA)
mRRA (ORA)

Figure 3. Evaluation of false discovery rate control for EPIC array data. (A) Type I error rates across 100 simulations for varying numbers of randomly sampled CpGs. **(B)** Median average numbers of CpGs per gene for GO categories with an unadjusted p-value < 0.05. The hypergeometric test is biased towards GO categories with more CpGs per gene on average. GOmeth = adjust for probe-number and multi-gene bias; HGT = hypergeometric test; HGT-mod = adjust for probe-number bias only. **(C)** Multidimensional scaling plot of normal samples from TCGA KIRC data, coloured by sex. **(D)** False discovery rate control of seven gene set testing methods. Two groups were generated by randomly sampling *n* samples per group, followed by differential methylation analysis and subsequent gene set testing. This was repeated 100 times at each sample size. The proportion of gene sets with unadjusted p-value < 0.05 across the 100 null simulations is shown for each method, at each sample size. Methods with good false discovery rate control should have relatively tight distributions around the red dashed line at 0.05. ebGSEA (KPMT) = ebGSEA using Known Population Median Test; ebGSEA (WT) = ebGSEA using Wilcoxon Test; GOmeth (1000) = GOmeth using top 1000 most significant probes; GOmeth (5000) = GOmeth using top 5000 most significant probes; mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.

as input. In order to compare the Type I error rate control of GOmeth with ebGSEA and *methylGSA* we analysed the normal samples from the TCGA 450K array kidney renal cell carcinoma (KIRC) dataset (Figure 3C). We took a resampling approach whereby we randomly assigned samples to one of two "groups" and varying the sample size per group (n = 5, 10, 20, 40, 80). We then performed differential methylation analysis between the two artificial groups, followed by gene set testing using the available methods: mGLM, mRRA (ORA), mRRA (GSEA), ebGSEA (WT), ebGSEA (KPMT), and GOmeth. We defined the input for GOmeth as either the top 1000 or top 5000 most highly ranked differentially methylated CpGs even though the probes did not reach statistical significance. This allowed us to calculate the proportions of gene sets that were significantly enriched for each of the methods, where "significant" is defined as a p-value less than 0.05. We repeated these steps 100 times at each sample size. In this scenario, where there are no true biological pathways differentially methylated, we expect 5% or fewer gene sets to be significantly enriched. Because ebGSEA can only test the specific gene sets available in the *ChAMP* package, which are based on gene sets from the Broad's Molecular Signatures Database (MSigDB), we limited all our comparisons to these 8567 gene sets. In addition, to ensure that the methods in the *methylGSA* package produced reasonable output, we limited the size of the gene sets to those with at least 5 and at most 5000 genes in the set. If these additional constraints are not included, the *methylGSA* methods mGLM and mRRA (ORA) produced results that were heavily biased towards reporting very small gene sets as highly significant. Furthermore, mRRA (ORA) was also biased towards ranking large gene sets very highly, if they were not filtered out (Supplementary Figure 3C).

In total, we compared seven different variants of the gene set tests: GOmeth with top 1000 CpGs, GOmeth with top 5000 CpGs, the three testing frameworks in *methylGSA* and the two tests in ebGSEA. In general, varying the sample size did not make a difference to the results, with consistent patterns observed for all sample sizes (Figure 3D). The worst performing test was mRRA (ORA), which had a median proportion of significantly enriched gene sets of at least 0.6 (or a median of 5388 enriched gene sets, across the 5 sample sizes). mRRA (GSEA) also performed poorly, with a median proportion of significantly enriched gene sets of at least 0.25 (or a median of 2117 enriched gene sets, across the 5 sample sizes). However, mGLM correctly controlled the false discovery rate at 0.05. The two variants of ebGSEA had only

slightly greater than 0.05 median proportion of false discoveries although their results were more variable across the 100 simulations at each sample size, with some simulations showing large numbers of false positives. GOmeth, using both top 1000 and 5000 CpGs, showed highly consistent performance, with median proportions of false discoveries < 0.05 (with a median of 180 and 278.5 enriched gene sets, respectively, across the 5 sample sizes), suggesting that GOmeth correctly controls for false discoveries.

Application to blood cell type EPIC data

Following our simulation studies, we wanted to test the performance of GOmeth, ebGSEA and the *methylGSA* methods on real data that contained differential methylation. We used a publicly available dataset of flow sorted blood cell types profiled on Illumina Infinium HumanMethylationEPIC arrays (GSE110554) (Salas et al. 2018). Cell types are easily distinguished based on methylation patterns (Figure 4A). We chose to perform our differential analysis and gene set testing on three independent pair-wise comparisons of cell types with varying numbers of differentially methylated probes: (1) CD4 vs CD8 T-cells, (2) monocytes vs neutrophils and (3) B-cells vs natural killer (NK) cells (Figure 4B). Differential methylation was performed using TREAT (McCarthy and Smyth 2009) and CpGs were defined as significantly differentially methylated if they had false discovery rates < 0.05 and $\Delta\beta$ cut-off of ~ 10% (corresponding to $\Delta M \sim 0.5$). Following the differential methylation analysis, we tested enrichment of GO sets and KEGG pathways using hypergeometric tests, GOmeth and the three *methylGSA* tests. Again, we limited the gene sets to those with a minimum of 5 and a maximum of 5000 genes for *methylGSA*. In order to compare to ebGSEA, we also tested for enrichment of the 8567 Broad gene sets using all gene set testing methods. For input to GOmeth, we used the top 5000 most highly ranked CpGs for each comparison, which is a subset of the total number of significant CpGs.

To evaluate the significant gene sets in a systematic way we took two approaches. First, we identified all immune categories in the GO database, as these are expected to be highly enriched when comparing different blood cell types. We therefore defined "true positive" GO categories as all the child terms under the parent GO category "immune system response" (GO:002376) from AMIGO 2 (http://amigo.geneontology.org/amigo/term/GO:0002376). We

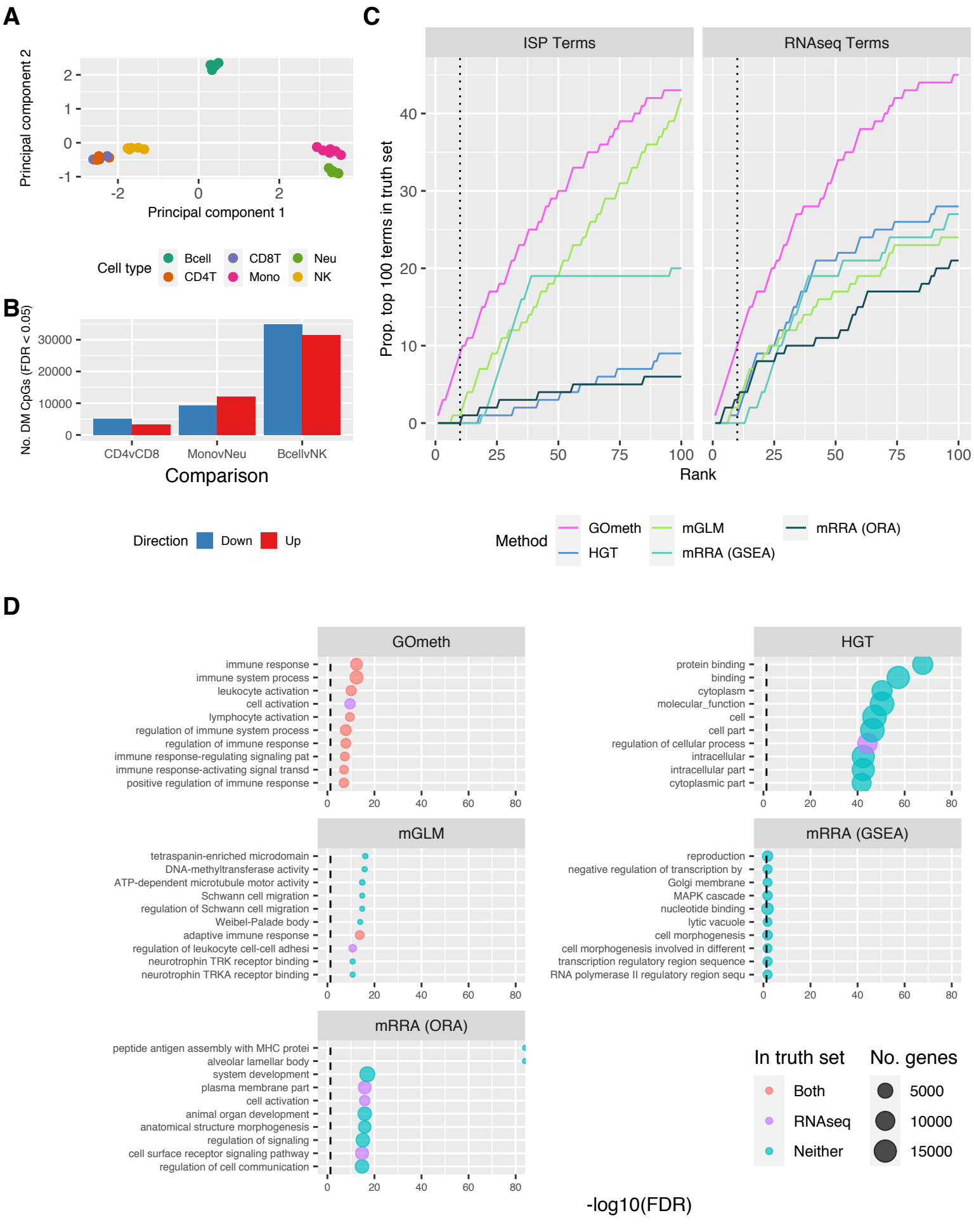


Figure 4. Comparison of gene set testing performance on Gene Ontology (GO) categories. **(A)** Multidimensional scaling plot of EPIC array sorted blood cell data. **(B)** Numbers of differentially methylated CpGs with an adjusted p-value < 0.05, for each cell type comparison: CD4 T-cells vs. CD8 T-cells, monocytes vs. neutrophils and B-cells vs. NK cells. The blue bar is the number of significant CpGs that are less methylated in the first cell type relative to the second, and the red bar is the number that are more methylated; e.g. ~5000 CpGs are less methylated and ~3000 are more methylated in CD4 T-cells, compared to CD8 T-cells. **(C)** Cumulative number of GO terms, as ranked by various methods, that are present in each truth set for the **B-cells vs. NK** comparison. ISP Terms = immune-system process child terms truth set; RNAseq Terms = top 100 terms from RNAseq analysis of the same cell types. **(D)** Bubble plots of the top 10 GO terms as ranked by various gene set testing methods. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the term is present in either RNAseq (purple) or ISP (green) truth sets, both (red) or neither (blue). GOmeth = GOmeth using top 5000 most significant probes; HGT = hypergeometric test; mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.

then counted how many of these immune sets were present in the top ranked gene sets for each method. We similarly defined true positives for the KEGG pathways by identifying all pathways belonging to the following categories: Immune system, Immune disease, Signal transduction, Signaling molecules and interaction (https://www.genome.jp/kegg/pathway.html). The second approach we took to evaluate the different methods was by analysing a publicly available RNA-Seq dataset comparing the same blood cell types (GSE107011; SRP125125) (Monaco et al. 2019; W. Xu et al. 2019). We performed differential expression analysis and gene set testing on the expression data and defined the top 100 significantly enriched gene sets from the RNA-Seq analysis as the "truth" (Supplementary Figure 4).

For GO categories, GOmeth always performed the best with the highest numbers of top ranked categories overlapping with "truth" sets across the three comparisons (Figure 4C-D, Supplementary figures 5-6). For KEGG pathways, the differences between the methods was not as clear, with mGLM and GOmeth generally the top performers except for the monocyte vs neutrophil comparison (Supplementary figure 7). In order to compare to ebGSEA we included the MSigDB sets and used the RNA-Seq gene sets as the "truth". Again, GOmeth and mGLM were top performers with the highest overlap of gene sets with the RNA-Seq gene sets (Supplementary figures 8-10).

Next we examined the top 10 ranked terms for each of the five gene set testing methods (Figure 4D, Supplementary figures 5-10). For the B-cells vs NK cells, we noted that the hypergeometric test tended to have very large, non-specific GO categories most highly ranked, with "protein binding", "cytoplasm" and "molecular function" in the top 10 (Figure 4D). The top 10 enriched GO categories for GOmeth were more biologically relevant, with immune specific gene set tests highly enriched (for example "leukocyte activation" and "lymphocyte activation"). All of the top 10 terms for GOmeth were included in at least one "truth" set. For the *methylGSA* methods, mGLM appeared to have more immune specific categories in the top 10 (e.g. "adaptive immune response" and " regulation of leukocyte cell-cell adhesion") compared to mRRA (GSEA) and mRRA (ORA). The results for mRRA (ORA) and mRRA (GSEA) were more difficult to interpret with none of the top 10 mRRA (GSEA) results included in either "truth" set. A similar pattern was observed for the CD4 vs CD8 T-cells and monocyte vs neutrophil comparisons, with GOmeth consistently ranking

more immune specific terms in the top 10 (Supplementary figures 5 and 6). We generally found, across all comparisons and different gene set ensembles, that the hypergeometric test, mRRA (ORA) and mRRA (GSEA) tended to have large, non-specific categories highly ranked, with GOmeth ranking more biologically relevant gene sets in the top 10.

Comparing compute time between gene set testing methods

While mGLM generally performs well, computationally, it was the slowest of all the methods to run on a single core (~50 minutes), with ebGSEA also taking in excess of 25 minutes to complete an analysis (Table 1). It is worth noting that the mGLM method can be parallelised to speed up computation, however, even using 9 cores it took approximately 8 minutes to complete the analysis of the MSigDB sets. By comparison, GOmeth is ~75 times faster than mGLM and ~41 times faster than ebGSEA. mRRA (ORA) is the fastest to run but generally does not perform as well as other methods.

Table 1: Average run-time across all contrasts. Gene sets used are Broad MSigDB gene sets from the ChAMP package. All methods were run on a single core.

Method	**Minutes**
mRRA (ORA)	0.16
GOmeth	0.67
mRRA (GSEA)	2.43
ebGSEA	27.72
mGLM	50.71

Gene set testing following a region based analysis

CpGs are not evenly spaced across the genome and often appear in clusters e.g. CpG islands (Gardiner-Garden and Frommer 1987); and several studies have demonstrated that CpGs in close proximity have correlated methylation levels (Eckhardt et al. 2006). Thus, rather than testing individual CpGs, identifying correlated methylation patterns between several spatially adjacent CpGs has been shown to yield more functionally relevant results (Hansen et al. 2011).

Several tools have been published for identifying differentially methylated regions (DMRs) from methylation array data: *Probe lasso* (Butcher and Beck 2015), *bumphunter* (Jaffe et al. 2012), *DMRcate* (Peters et al. 2015), *mCSEA* (Martorell-Marugán, González-Rumayor, and Carmona-Sáez 2019) and *DMRforPairs* (Rijlaarsdam et al. 2014). Depending on the input data, region finding tools can identify several hundred or even thousands of DMRs. They all generally output the location of the region including the chromosome, region start and region end positions, along with some additional metrics and statistical significance. Some tools attempt to annotate the regions with genes but others do not. Thus, when faced with a long list of DMRs it is unclear how to interpret the biological significance of the results, and there are no gene set testing tools available for DMRs.

To address this, we have developed GOregion; an extension of GOmeth, that enables gene set testing of DMRs. The "goregion" function tests GO terms and KEGG pathways, whilst "gsaregion" is a generalised function that accepts any list of gene sets as input. We reasoned that because region detection is inherently dependent on CpG probe density, DMRs are more likely to be identified in genes with more CpG probes. This trend is observed in the blood cell type data (Figure 5A, Supplementary figure 11A, 12A). To take this bias into account, GOregion utilises the GOmeth testing framework. GOregion accepts a ranged object of DMRs that have been identified by the user's choice of region-finding software. These regions are then overlapped with the locations of the CpGs on the Illumina array to identify a set of CpGs underlying the DMRs. These CpG probes are then passed to GOmeth and GOmeth's existing algorithm is used to test for enrichment of gene sets.

Using the sorted blood cell data we identified DMRs for the same three cell type comparisons using the *DMRcate* package: 1) CD4 vs CD8 T-cells, 2) monocytes vs neutrophils and 3) B-cells vs. NK cells. Using default parameters, *DMRcate* identified 6404, 7176 and 23,210

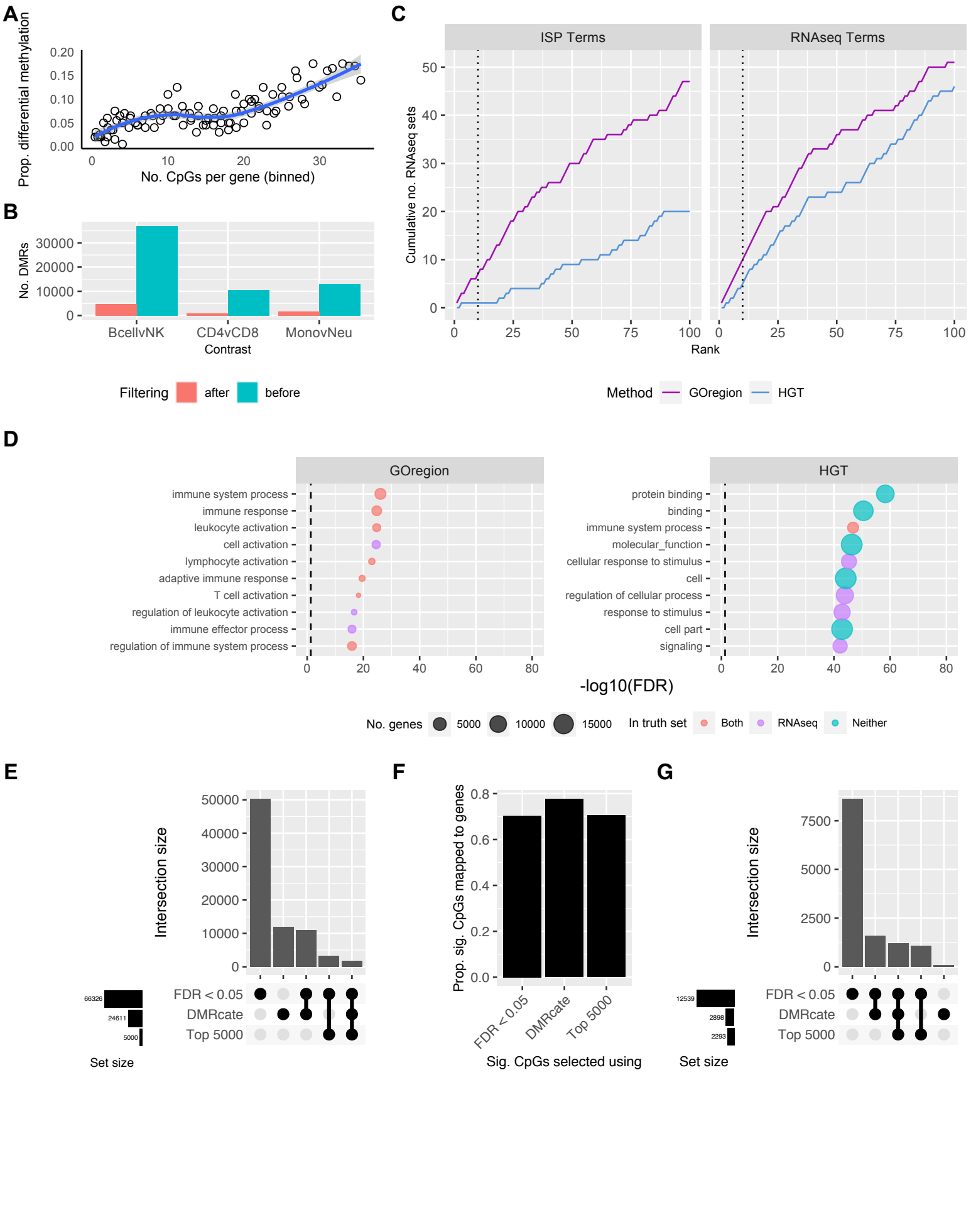


Figure 5. Evaluation of the performance of GOregion on sorted blood cell data. (A) Bias plot showing that genes with more measured CpGs are more likely to have a differentially methylated region (DMR). This plot is produced from EPIC array sorted blood cell type data, comparing B-cells to NK cells. **(B)** Numbers of DMRs identified by *DMRcate*, for each cell type comparison: CD4 T-cells vs. CD8 T-cells, monocytes vs. neutrophils and B-cells vs. NK cells. The blue bar is the number of DMRs before filtering, the pink bar is the number of DMRs after filtering out DMRs with < 3 underlying CpGs and an absolute mean $|\Delta\beta| < 0.1$. **(C)** Cumulative number of GO terms, as ranked by GOregion and a simple hypergeometric test (HGT), that are present in each truth set for the B-cells vs. NK comparison. ISP Terms = immune-system process child terms truth set; RNAseq Terms = top 100 terms from RNAseq analysis of the same cell types. **(D)** Bubble plots of the top 10 GO terms as ranked by GOregion and a simple HGT for the B-cells vs. NK comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the term is present in either RNAseq (purple) or ISP (green) truth sets, both (red) or neither (blue). **(E)** Upset plot showing the characteristics of the CpGs selected as "significant" for the B-cell vs. NK comparison by a probe-wise differential methylation analysis using a significance cut off (FDR < 0.05), the top 5000 CpGs as ranked by the probe-wise analysis (Top 5000) or the CpGs underlying the filtered *DMRcate* regions (DMRcate). The probe-wise analysis with FDR < 0.05 identified over 60,000 CpGs as "significant" and had the most unique CpGs. However, despite identifying fewer "significant" CpGs (~25,000), almost half of the CpGs identified by DMRcate are unique (~12,000). **(F)** Proportion of "significant" CpGs that are annotated to genes as identified by the three different strategies. **(G)** Upset plot showing the characteristics of the genes that "significant" CpGs are annotated to, as identified by the three different strategies, for the B-cell vs. NK comparison. CpGs identified by the probe-wise analysis with FDR < 0.05 map to over 12,000 genes. Although the CpGs identified by DMRcate map to far fewer genes (~2900), a number of them are unique to this approach.

differentially methylated regions, respectively. We further filtered the DMRs by only including regions containing 3 or more CpGs with an absolute mean β value difference greater than 0.1. This left 789, 1633 and 4723 DMRs, respectively, for downstream analysis (Figure 5B). We then performed gene set testing of GO categories using GOregion, and compared it to a simple approach of overlapping DMRs with known genes, and then testing using a HGT.

As previously described, we evaluated the results by counting the numbers of highly ranked immune-related GO terms, and the numbers of highly ranked GO categories identified in the RNA-Seq data analysis of the same cell types. GOregion consistently ranked immune-related and RNA-Seq "truth" terms more highly than the simple HGT-based strategy (Figure 5C, Supplementary figures 11B, 12B). Examining the top 10 most highly ranked GO categories showed that GOregion categories are more specific to immune processes than those identified using the HGT approach (Figure 5D, Supplementary figure 11C, 12C). Across all comparisons, GOregion always ranked more "truth" sets in the top 10 than HGT. For example, the top 10 gene sets ranked by GOregion for the B-cells vs NK comparison were highly specific to immune system processes, e.g. "T cell activation", whereas for HGT the sets were very broad and contained thousands of genes e.g. "protein binding" (Figure 5D).

We wanted to explore the differences between a region analysis and a probe-wise analysis and how this would affect gene set testing results. GOmeth is highly dependent on the set of significant CpGs provided as input, so we postulated that selecting CpGs using a region-level analysis could be more biologically relevant, in certain circumstances. We compared the numbers of differentially methylated probes and genes that are selected based on a region analysis with *DMRCate*, a probe-wise analysis using an FDR cut-off < 0.05, and a probe-wise analysis selecting the top 5000 differentially methylated CpGs. For the B-cell vs NK cells comparison, a probe-wise analysis with FDR < 0.05 selects over 60,000 differentially methylated CpGs. More than 50,000 of these CpGs are unique to this approach and not identified with the region approach, whereas *DMRCate* identifies ~24,000 CpGs in DMRs (Figure 5E). We noted that although *DMRCate* always captured fewer significant CpGs (Figure 5E, Supplementary figure 11D, 12D), a higher proportion of the total significant CpGs were annotated to genes compared with either of the probe-wise approaches (Figure 5F, Supplementary Figures 11E, 12E). This results in sets of genes uniquely captured using a

region-based analysis (Figure 5G, Supplementary Figures 11F, 12F).

Comparing GOregion to the two probe-wise approaches using our previously defined "truth" sets showed that all of the approaches performed similarly well across all the contrasts, except for B-cells vs NK cells (Supplementary figure 13A, B). For that comparison, GOregion and the GOmeth probe-wise analysis using the top 5000 CpGs both performed markedly better than GOmeth with significant CpGs selected using FDR < 0.05 (Supplementary figure 13A). This is likely due to these significant CpGs being annotated to > 12,000 genes, resulting in a highly non-specific set of genes as input to GOmeth, whereas *DMRCate* regions are annotated to just under 3000 genes (Figure 5G). Examining the terms that were highly ranked by the various approaches, across the different contrasts, revealed that they all tended to be immune specific (Supplementary figure 13C-E). The exception is the B-cells vs NK comparison produced by GOmeth based on CpGs selected at FDR < 0.05, where the most highly ranked terms were very broad categories such as "cell communication", "signalling" and "plasma membrane part" (Supplementary figure 13C). Hence, for comparisons which result in very large numbers of significantly differentially methylated CpGs, limiting the set of CpGs used as input for gene set testing, such as performing a region analysis, is important for producing meaningful results.

Region-finding software is itself dependent on numerous parameters and appropriate downstream filtering of results can also be important in identifying the most biologically informative set of DMRs. Using the blood cell type dataset, we demonstrate that different DMR filters affect the downstream GOregion gene set testing results in different ways, depending on the comparison (Supplementary Figure 14). For our dataset, the most important filter to include is a mean $\Delta\beta$ difference cut-off, i.e. the size of the methylation difference. For this particular dataset, we find that a $\Delta\beta$ difference cut-off of 0.1 performs best across all comparisons. For the B-cells vs NK cells comparison, we noticed that not filtering with a $\Delta\beta$ cut-off produced very poor results. In contrast, for the monocyte vs neutrophils comparison, a $\Delta\beta$ cut-off of 0.2 was too stringent and resulted in fewer gene sets overlapping with the "truth" sets (Supplementary figure 14A, B).

GOregion is compatible with the results of any software for finding differentially methylated regions which can be expressed as a ranged data object. It is also computationally efficient

and can test a variety of gene sets such as GO categories, KEGG pathways or any list of custom gene sets.

Discussion

Gene set testing is a useful tool to gain additional biological insight into the underlying mechanisms in an experiment. Here we present GOmeth for performing gene set testing after a probe-level analysis, and GOregion, a gene set testing method following a region-based analysis. To the best of our knowledge, GOregion is the only method that specifically tests enrichment of gene sets for differentially methylated regions. Both of these methods take into account probe-number and multi-gene bias for the analysis of 450K and EPIC Illumina HumanMethylation arrays.

Through the use of simulations and resampling normal samples we have shown that GOmeth correctly controls the false discovery rate with minimal bias. We applied GOmeth to a blood cell type dataset and showed that the top ranked categories are consistently biologically relevant across multiple cell type comparisons. We defined two different types of "truth" sets based on the information available in the GO and KEGG databases, as well as an independently analysed RNA-Seq dataset. We acknowledge that our "truth" sets will have shortcomings in that they are unlikely to encompass all the truly enriched pathways, and, particularly in the case of the RNA-Seq data analysis, the choice of gene set testing method is likely to play a role in how gene sets are ranked. Nevertheless, we showed that GOmeth generally outperforms other available methods. Further, we have shown that the probe-number bias affects the probability that a region is called differentially methylated. We therefore developed GOregion to perform unbiased gene set testing following a region-based analysis. For the blood cell type dataset, GOregion outperformed a simple hypergeometric testing approach that has been previously used in analyses.

An important consideration when using GOmeth is how many differentially methylated probes to use as input to the method. For comparisons that have tens of thousands of significantly differentially methylated CpGs, in the first instance, we would recommend performing a region-based analysis and using GOregion to perform gene set testing. Another

option would be to use GOmeth, but restrict the input CpGs to the top ranked CpGs, with a rule of thumb that the number of input CpGs is less than 10,000. Simply taking a false discovery rate cut-off for the B-cells vs NK cells comparison, for example, leads to too many genes being identified as differentially methylated (>12,000) and we found that the gene set testing results were not very specific or biologically meaningful. In our comparisons, we have used the top 5000 differentially methylated probes and this produced good results for the blood cell type dataset. There is additional functionality in GOmeth to restrict the list of significant CpGs by genomic features, such as "TSS1500", "TSS200" and "Body", for example. This has the effect of decreasing the overall numbers of CpGs to use as input, but potentially retains more biologically meaningful loci.

Even though a region-based approach can potentially select fewer CpGs, in our analysis of the blood cell data, it always identified numerous unique CpGs not detected using the probe-wise methods. This is likely because these CpGs were not statistically significant on their own but are identified as part of a region. Thus, given the potential for capturing biologically important CpGs, which may be missed by probe-wise approaches due to their reliance on rankings and significance cut-offs, we suggest that a good quality, region-based analysis can potentially distil more focused gene sets than a probe-level gene set analysis of the same data.

A very important feature of GOmeth and GOregion is their flexibility. Unlike the *methylGSA* methods, GOmeth and GOregion do not require any filtering of gene sets to produce robust results. Furthermore, in contrast to ebGSEA, GOmeth and GOregion can perform gene set enrichment analysis using a variety of gene sets; including GO categories, KEGG pathways or any other list of gene sets supplied by the user.

Conclusions

GOmeth and GOregion are novel statistical methods to perform unbiased gene set testing for methylation arrays. We have shown that our methods produce the most biologically meaningful results while controlling the false discovery rate. All of our gene set testing functions are available in the *missMethyl* Bioconductor R package.

Methods

All analysis code presented in this manuscript can be found at http://oshlacklab.com/methyl-geneset-testing/. The analysis website was created using the *workflowr* (1.6.2) R package (Blischak, Carbonetto, and Stephens 2019). The GitHub repository associated with the analysis website is at: https://github.com/Oshlack/methyl-geneset-testing.

Statistical model for GOmeth

The statistical test for GOmeth and GOregion is based on Wallenius' noncentral hypergeometric distribution, which is a generalised version of the hypergeometric distribution where items are sampled with bias. For GOmeth, we take the following stepwise procedure:

1. For each CpG i annotated to gene j, calculate a weight

$$w_{ij} = \frac{1}{\#genes\ annotated\ to\ CpG_{ij}}$$

2. Let $i = 1, \ldots, I_j$ denote the CpGs annotated to gene j. Calculate the equivalent number of CpGs measured across gene j as:

$$N_j = \sum_{i=1}^{I_j} w_{ij}$$

3. Let A define the set of significant differentially methylated Cpgs. For each gene j, define an indicator vector $\mathbf{1}_j(x)$ of length I_j such that $x_i = 1$ if $CpG_{ij} \in A$, and $x_i = 0$ if $CpG_{ij} \notin A$, where $i = 1, \ldots, I_j$.

4. Let $\boldsymbol{w}_j$ define the vector of weights w_{ij} for each gene j. Calculate the differential methylation score for each gene j

$$S_j = \min(\mathbf{1}_j(x) \cdot \boldsymbol{w}_j, 1)$$

5. Let $j = 1, \ldots, J_g$ denote the genes that are present in gene set g. Calculate the enrichment statistic for Wallenius' noncentral hypergeometric test for each gene set g

$$ES_g = \sum_{j=1}^{J_g} S_j$$

6. Calculate the probability weighting function (PWF) by applying a moving average smoother to an ordered binary vector (based on the number of associated CpGs) where 1 indicates a gene is differentially methylated and 0 indicates the gene is not differentially methylated. We use the 'tricubeMovingAverage' function in the *limma* package which is similar to a least squares loess curve of degree zero. The binary vector is ordered by the number of CpGs measuring methylation across each gene, N_j, from smallest to largest. The output is a vector of the same length as the input such that each gene is assigned a probability of differential methylation based on the smoothed value. We then calculate the expected odds of enrichment for each gene set g by calculating the mean PWF of the genes in the set and comparing it to the mean PWF of the rest of the genes represented on the array.

$$ODDS_g = \frac{mean(PWF(genes\ in\ gene\ set\ g))}{mean(PWF(genes\ not\ in\ gene\ set\ g))}$$

7. For testing enrichment of each gene set g, we obtain a one-sided p-value from Wallenius' noncentral hypergeometric distribution with the following parameters: $x = \lfloor ES_g \rfloor$, m_1 = the size of the gene set J_g , m_2 = the number of genes on the rest of the array, n = the total number of significant genes, and $odds = ODDS_g$. We use the *BiasedUrn* R package to obtain p-values.

Null simulations: random sampling of CpGs

We randomly selected sets of 50, 100, 500, 1000, 5000 and 10,000 CpGs from the Illumina array annotation for both 450k and EPIC arrays. The sampling was repeated 100 times for each CpG set size. We tested for significant enrichment of GO categories using a standard hypergeometric test (HGT), a Wallenius' hypergeometric test accounting for probe number

bias (HGT-mod) and GOmeth, which is based on Wallenius' hypergeometric test and accounts for probe number and multi-gene bias.

Methylation Datasets

The normal samples from the KIRC TCGA dataset (Cancer Genome Atlas Research Network 2013) were used for estimating the false discovery rate of the different gene set testing methods. The data was downloaded using the *curatedTCGAData* Bioconductor package (Ramos 2020) and the 160 normal samples extracted. The data was provided as already processed β values, however we performed additional filtering and removed poor quality probes, probes containing SNPs as well as sex chromosome probes. The resulting multidimenational scaling plots showed no apparent evidence of sex or other technical effects (Figure 3C).

To compare the performance between different gene set testing methods when there is significant differential methylation, we used Illumina Infinium HumanMethylationEPIC (GSE110554) data generated from flow-sorted neutrophils (Neu, n = 6), monocytes (Mono, n = 6), B-lymphocytes (Bcells, n = 6), CD4+ T-cells (CD4T, n=7, six samples and one technical replicate), CD8+ T-cells (CD8T, n = 6), Natural Killer cells (NK, n = 6) and 12 DNA artificial mixtures (labeled as MIX) (Salas et al. 2018). Only the sorted cells were used in our analysis. The data was downloaded using the *ExperimentHub* (1.12.0) Bioconductor package.

Quality control and normalization

Analysis was performed using R (3.6.3) (R Core Team 2014). All data was processed using the *minfi* (1.32.0) (Aryee et al. 2014; Fortin, Triche, and Hansen 2017) R Bioconductor (R. C. Gentleman et al. 2004; Huber et al. 2015) package. Between array and probe-type normalization was performed using the stratified quantile normalisation (SQN) method (Touleimat and Tost 2012). Probes with a detection P-value > 0.01 in one or more samples were discarded. Probes potentially affected by common SNPs (minor allele frequency > 0)

proximal to the CpG of interest (up to 2 bp upstream and 1 downstream) and non-specific probes (Pidsley et al. 2016; Y.-A. Chen et al. 2013) were also removed from further analysis.

Statistical analysis

The proportion of methylation at each CpG is represented by the β value, defined as the proportion of the methylated signal to the total signal and calculated from the normalized intensity values. Statistical analyses were performed on M-values $\left[M = \frac{methylated}{unmethylated}\right]$ as recommended by Du et al. (2010).

Comparison of gene set testing methods using sorted blood cell data

CpG probe-wise linear models were fitted to determine differences in methylation between cell types (B-cells vs NK, CD4 vs CD8 T-cells, monocytes vs neutrophils) using the *limma* (3.42.2) package (Ritchie et al. 2015). Differentially methylated probes (DMPs) were identified using empirical Bayes moderated t-tests (Smyth 2005), performing robust empirical Bayes shrinkage of the gene-wise variances to protect against hypervariable probes (Phipson et al. 2016). Empirical Bayes moderated-t p-values were then calculated relative to a minimum meaningful log-fold-change (lfc) threshold on the M-value scale (lfc = 0.5, corresponding to $|\Delta\beta| \sim 0.1$) (McCarthy and Smyth 2009). P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995).

For each comparison, we tested for significant enrichment of GO categories and KEGG pathways. We only tested sets with at least 5 genes and at most 5000 genes for the *methylGSA* methods. The top ranked 5000 CpGs were tested using the HGT and GOmeth, from the *missMethyl* (1.20.4) package, for enrichment of GO terms and KEGG pathways. The raw p-values were passed as input to the *methylGSA* (1.4.9) methods.

The ebGSEA method only tests a built-in set of Broad MSigDB gene sets for enrichment. To compare other methods with ebGSEA, all methods were tested on the gene sets supplied with

the *ChAMP* (2.16.2) package. The gene sets were restricted to sets with at least 5 genes or at most 5000 for the *methylGSA* methods.

Comparison of gene set testing methods using kidney clear cell carcinoma (KIRC) data

The KIRC data from the *curatedTCGAData* (1.8.1) package was provided as β values with masked data points; data points were masked as "NA" if their detection p-value was greater than 0.05 or the probe was annotated as having a SNP within 10 base pairs or repeat within 15 base pairs of the interrogated CpG (Cancer Genome Atlas Research Network et al. 2016). We extracted only the 160 normal samples and removed probes with any "NA" values, as well as SNP-affected probes and multi-mapping and sex-chromosome probes, as previously described. This left 364,602 probes for downstream analysis.

We ran 100 null simulations by randomly subsampling the normal samples and splitting them into two artificial "groups" with 5, 10, 20, 40 and 80 samples per "group". For each of the 100 simulations, at each sample size, DMPs between "groups" were identified using empirical Bayes moderated t-tests (Smyth 2005), performing robust empirical Bayes shrinkage of the gene-wise variances to protect against hypervariable probes (Phipson et al. 2016).

We then performed gene set testing of the differential methylation analysis results using several methods with the Broad MSigDB gene sets available in the *ChAMP* Bioconductor package. GOmeth was run using both the top 1000 and top 5000 significant CpGs as input. The *methylGSA* methods; mGLM, mRRA (ORA) and mRRA (GSEA) were run with gene set sizes restricted to a minimum of 5 and maximum of 5000 genes. The ebGSEA method was run using default parameters and both its KPMT and WT output were compared.

RNA-Seq data and analysis

The RNA-Seq data for the sorted blood cell types was downloaded from SRA (GSE107011; SRP125125) (Monaco et al. 2019; W. Xu et al. 2019). The reads were mapped to hg19 reference transcriptome

(http://refgenomes.databio.org/v2/asset/hg19_cdna/fasta/archive?tag=default) and quantified using Salmon (1.2.1) (Patro et al. 2017). Salmon transcript-level estimates were imported and summarised at the gene-level as length-scaled TPM using the *tximport* (1.14.2) Bioconductor package (Soneson, Love, and Robinson 2015). Lowly expressed genes were filtered out using the *edgeR* (3.28.1) (Robinson, McCarthy, and Smyth 2010) 'filterByExpr' function as described by Chen at al. (2016). The data was then TMM normalised (Robinson and Oshlack 2010) and transformed using 'voomWithQualityWeights' (Liu et al. 2015), to increase power by combining 'voom' (Law et al. 2014) observational-level weights with sample-specific weights.

Probe-wise linear models were then fitted for each gene to determine gene expression differences between cell types (B-cells vs NK cells, CD4 vs CD8 T-cells, monocytes vs neutrophils) using *limma* (3.42.2) (Ritchie et al. 2015). Differentially expressed genes were identified using empirical Bayes moderated t-tests (Smyth 2005), performing robust empirical Bayes shrinkage of the gene-wise variances to protect against hypervariable probes (Phipson et al. 2016). P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995).

We used the 'goana' function from the *limma* (3.42.2) package to test enrichment of GO categories, 'kegga' to test for enrichment of KEGG pathways and a generalised version of 'goana' and 'kegga'; to test for enrichment of the Broad MSigDB gene sets. All the methods took gene length bias into account (Young et al. 2010). GO, KEGG and MSigDB "truth" sets were then defined for each cell type comparison from the RNA-Seq analysis as the top 100 enriched sets.

Evaluation of GOregion using flow sorted blood cell data

The lllumina Infinium HumanMethylationEPIC (GSE110554) data generated from flow-sorted blood cells was used for identification of DMRs. The data was processed as previously described. DMRs between cell types (B-cells vs NK cells, CD4 vs CD8 T-cells, monocytes vs neutrophils) were identified using the *DMRcate* (2.0.7) Bioconductor package (Peters et al. 2015). The analysis was performed on M-values using default parameters.

Downstream gene set testing was performed on a filtered list of DMRs with a mean $|\Delta\beta| \geq 0.1$ and at least 3 underlying CpGs.

GO terms were tested for enrichment of DMR-associated genes using 'goregion' and a standard HGT, as implemented in the 'goana' function from the *limma* (3.42.2) Bioconductor package. A gene, as defined in the *TxDb.Hsapiens.UCSC.hg19.knownGene* Bioconductor package (3.2.2), was included in the list of genes to be tested using 'goana' if it overlapped a DMR by at least 1bp.

Abbreviations

TCGA: The Cancer Genome Atlas

MSigDB: Molecular Signatures Database

RNA-Seq: RNA sequencing

FDR: false discovery rate

DMR: differentially methylated region

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data used in this manuscript is publicly available. The KIRC data was downloaded using the *curatedTCGAData* (1.8.1) R Bioconductor package. The sorted blood cell type Illumina Infinium HumanMethylationEPIC data is available from the Gene Expression Omnibus under accession number GSE110554 and was downloaded using the *ExperimentHub* (1.12.0) R Bioconductor package. The RNA-seq data for the sorted blood cell types was downloaded from SRA (GSE107011; SRP125125). All analysis code presented in this manuscript can be found at the following *workflowr* website http://oshlacklab.com/methyl-geneset-testing/. The GitHub repository associated with the analysis website is at: https://github.com/Oshlack/methyl-geneset-testing.

Competing interests

The authors declare that they have no competing interests.

Funding

BP is supported by an Emerging Leader Investigator Grant (GNT1175653) from the National Health and Medical Research Council (NHMRC). AO is supported by NHMRC GNT1126157. The funding body did not play any role in the study design, analysis, interpretation of data or writing of the manuscript.

Authors' contributions

The original idea for gene set testing for methylation array data was conceived by AO. BP conceived the statistical model and wrote the core of the 'gometh' and 'gsameth' code. JM conceived the idea of extending GOmeth to DMRs. BP and JM determined how to apply GOmeth to DMRs. JM wrote the GOregion code. JM performed all analysis and produced the figures and analysis website. BP and JM wrote the initial draft of the manuscript, which was revised by all authors. All authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge Peter Langfelder for providing detailed code and simulations which helped us to discover and correct for multi-gene bias in our testing framework. We would also like to acknowledge all *missMethyl* users who have provided valuable bug reports and feedback.

References

Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. "Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays." *Bioinformatics* 30 (10): 1363–69.

Assenov, Yassen, Fabian Müller, Pavlo Lutsik, Jörn Walter, Thomas Lengauer, and Christoph Bock. 2014. "Comprehensive Analysis of DNA Methylation Data with RnBeads." *Nature Methods* 11 (11): 1138–40.

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–500.

Blischak, John D., Peter Carbonetto, and Matthew Stephens. 2019. "Creating and Sharing Reproducible Research Code the Workflowr Way." *F1000Research* 8 (October): 1749.

Butcher, Lee M., and Stephan Beck. 2015. "Probe Lasso: A Novel Method to Rope in Differentially Methylated Regions with 450K DNA Methylation Data." *Methods* 72 (January): 21–28.

Cancer Genome Atlas Research Network. 2013. "Comprehensive Molecular Characterization of Clear Cell Renal Cell Carcinoma." *Nature* 499 (7456): 43–49.

Cancer Genome Atlas Research Network, W. Marston Linehan, Paul T. Spellman, Christopher J. Ricketts, Chad J. Creighton, Suzanne S. Fei, Caleb Davis, et al. 2016. "Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma." *The New England Journal of Medicine* 374 (2): 135–45.

Chen, Yi-An, Mathieu Lemire, Sanaa Choufani, Darci T. Butcher, Daria Grafodatskaya, Brent W. Zanke, Steven Gallinger, Thomas J. Hudson, and Rosanna Weksberg. 2013. "Discovery of Cross-Reactive Probes and Polymorphic CpGs in the Illumina Infinium HumanMethylation450 Microarray." *Epigenetics: Official Journal of the DNA Methylation Society* 8 (2): 203–9.

Chen, Yunshun, Aaron T. L. Lun, and Gordon K. Smyth. 2016. "From Reads to Genes to Pathways: Differential Expression Analysis of RNA-Seq Experiments Using Rsubread and the edgeR Quasi-Likelihood Pipeline." *F1000Research* 5 (June): 1438.

Davis, Sean, Pan Du, Sven Bilke, Tim Triche Jr., and Moiz Bootwalla. 2019. "Methylumi: Handle Illumina Methylation Data."

Dong, Danyue, Yuan Tian, Shijie C. Zheng, and Andrew E. Teschendorff. 2019. "ebGSEA: An Improved Gene Set Enrichment Analysis Method for Epigenome-Wide-Association Studies." *Bioinformatics* , January. https://doi.org/10.1093/bioinformatics/btz073.

Du, Pan, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren a. Kibbe, Lifang Hou, and Simon M. Lin. 2010. "Comparison of Beta-Value and M-Value Methods for Quantifying Methylation Levels by Microarray Analysis." *BMC Bioinformatics* 11 (1): 587.

Eckhardt, Florian, Joern Lewin, Rene Cortese, Vardhman K. Rakyan, John Attwood, Matthias Burger, John Burton, et al. 2006. "DNA Methylation Profiling of Human Chromosomes 6, 20 and 22." *Nature Genetics* 38 (12): 1378–85.

Esteller, Manel. 2007. "Cancer Epigenomics: DNA Methylomes and Histone-Modification Maps." *Nature Reviews. Genetics* 8 (4): 286–98.

Fortin, Jean-Philippe, Timothy J. Triche Jr, and Kasper D. Hansen. 2017. "Preprocessing, Normalization and Integration of the Illumina HumanMethylationEPIC Array with Minfi." *Bioinformatics* 33 (4): 558–60.

Gardiner-Garden, M., and M. Frommer. 1987. "CpG Islands in Vertebrate Genomes." *Journal of Molecular Biology* 196 (2): 261–82.

Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling,

Sandrine Dudoit, Byron Ellis, et al. 2004. "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology* 5 (10): R80.

Hansen, Kasper Daniel, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G. McDonald, Bo Wen, et al. 2011. "Increased Methylation Variation in Epigenetic Domains across Cancer Types." *Nature Genetics* 43 (8): 768–75.

Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21.

Jaffe, Andrew E., Peter Murakami, Hwajin Lee, Jeffrey T. Leek, M. Daniele Fallin, Andrew P. Feinberg, and Rafael A. Irizarry. 2012. "Bump Hunting to Identify Differentially Methylated Regions in Epigenetic Epidemiology Studies." *International Journal of Epidemiology* 41 (1): 200–209.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2): R29.

Liu, Ruijie, Aliaksei Z. Holik, Shian Su, Natasha Jansz, Kelan Chen, Huei San Leong, Marnie E. Blewitt, Marie-Liesse Asselin-Labat, Gordon K. Smyth, and Matthew E. Ritchie. 2015. "Why Weight? Modelling Sample and Observational Level Variability Improves Power in RNA-Seq Analyses." *Nucleic Acids Research* 43 (15): e97.

Martorell-Marugán, Jordi, Víctor González-Rumayor, and Pedro Carmona-Sáez. 2019. "mCSEA: Detecting Subtle Differentially Methylated Regions." *Bioinformatics* 35 (18): 3257–62.

McCarthy, Davis J., and Gordon K. Smyth. 2009. "Testing Significance Relative to a Fold-Change Threshold Is a TREAT." *Bioinformatics* 25 (6): 765–71.

Mi, Gu, Yanming Di, Sarah Emerson, Jason S. Cumbie, and Jeff H. Chang. 2012. "Length Bias Correction in Gene Ontology Enrichment Analysis Using Logistic Regression." *PloS One* 7 (10): e46128.

Monaco, Gianni, Bernett Lee, Weili Xu, Seri Mustafah, You Yi Hwang, Christophe Carré, Nicolas Burdin, et al. 2019. "RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types." *Cell Reports* 26 (6): 1627–40.e7.

Morris, Tiffany J., Lee M. Butcher, Andrew Feber, Andrew E. Teschendorff, Ankur R. Chakravarthy, Tomasz K. Wojdacz, and Stephan Beck. 2014. "ChAMP: 450k Chip Analysis Methylation Pipeline." *Bioinformatics* 30 (3): 428–30.

Müller, Fabian, Michael Scherer, Yassen Assenov, Pavlo Lutsik, Jörn Walter, Thomas Lengauer, and Christoph Bock. 2019. "RnBeads 2.0: Comprehensive Analysis of DNA Methylation Data." *Genome Biology* 20 (1): 55.

Oytam, Yalchin, Fariborz Sobhanmanesh, Konsta Duesing, Joshua C. Bowden, Megan Osmond-McLeod, and Jason Ross. 2016. "Risk-Conscious Correction of Batch Effects: Maximising Information Extraction from High-Throughput Genomic Datasets." *BMC Bioinformatics* 17 (1): 332.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19.

Peters, Timothy J., Michael J. Buckley, Aaron L. Statham, Ruth Pidsley, Katherine Samaras, Reginald V Lord, Susan J. Clark, and Peter L. Molloy. 2015. "De Novo Identification of Differentially Methylated Regions in the Human Genome." *Epigenetics & Chromatin* 8 (1): 6.

Phipson, Belinda, Stanley Lee, Ian J. Majewski, Warren S. Alexander, and Gordon K. Smyth. 2016. "ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION." *The Annals of Applied Statistics* 10 (2): 946–63.

Phipson, Belinda, Jovana Maksimovic, and Alicia Oshlack. 2016. "missMethyl: An R Package for Analyzing Data from Illumina's HumanMethylation450 Platform." *Bioinformatics* 32 (2): 286–88.

Phipson, Belinda, and Alicia Oshlack. 2014. "DiffVar: A New Method for Detecting Differential Variability with Application to Methylation in Cancer and Aging." *Genome Biology* 15 (9): 465.

Pidsley, Ruth, Chloe C. Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, and Leonard C. Schalkwyk. 2013. "A Data-Driven Approach to Preprocessing Illumina 450K Methylation Array Data." *BMC Genomics* 14 (1): 293.

Pidsley, Ruth, Elena Zotenko, Timothy J. Peters, Mitchell G. Lawrence, Gail P. Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J. Clark. 2016. "Critical Evaluation of the Illumina MethylationEPIC BeadChip Microarray for Whole-Genome DNA Methylation Profiling." *Genome Biology* 17 (1): 208.

Ramos, Marcel. 2020. "curatedTCGAData: Curated Data From The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects."

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org/.

Ren, Xu, and Pei Fen Kuan. 2018. "methylGSA: A Bioconductor Package and Shiny App for DNA Methylation Data Length Bias Adjustment in Gene Set Testing." *Bioinformatics* , October. https://doi.org/10.1093/bioinformatics/bty892.

Rijlaarsdam, Martin A., Yvonne G. van der Zwan, Lambert C. J. Dorssers, and Leendert H. J. Looijenga. 2014. "DMRforPairs: Identifying Differentially Methylated Regions between Unique Samples Using Array Based Methylation Profiles." *BMC Bioinformatics* 15 (1): 141.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.

Robinson, Mark D., and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11 (3): R25.

Salas, Lucas A., Devin C. Koestler, Rondi A. Butler, Helen M. Hansen, John K. Wiencke, Karl T. Kelsey, and Brock C. Christensen. 2018. "An Optimized Library for Reference-Based Deconvolution of Whole-Blood Biospecimens Assayed Using the Illumina HumanMethylationEPIC BeadArray." *Genome Biology* 19 (1): 64.

Smyth, G. K. 2005. "Limma: Linear Models for Microarray Data." In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, edited by Robert Gentleman, Vincent J. Carey, Wolfgang Huber, Rafael A. Irizarry, and Sandrine Dudoit, 397–420. New York, NY: Springer New York.

Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2015. "Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences." *F1000Research* 4 (December): 1521.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Touleimat, Nizar, and Jörg Tost. 2012. "Complete Pipeline for Infinium(®) Human Methylation 450K BeadChip Data Processing Using Subset Quantile Normalization for Accurate DNA Methylation Estimation." *Epigenomics* 4 (3): 325–41.

Wu, Di, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E. Visvader, and Gordon K. Smyth. 2010. "ROAST: Rotation Gene Set Tests for Complex Microarray Experiments." *Bioinformatics* 26 (17): 2176–82.

Wu, Di, and Gordon K. Smyth. 2012. "Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation." *Nucleic Acids Research* 40 (17): e133.

Xu, Weili, Gianni Monaco, Eleanor Huijin Wong, Wilson Lek Wen Tan, Hassen Kared, Yannick Simoni, Shu Wen Tan, et al. 2019. "Mapping of γ/δ T Cells Reveals Vδ2+ T

Cells Resistance to Senescence." *EBioMedicine* 39 (January): 44–58.

Xu, Zongli, Liang Niu, Leping Li, and Jack A. Taylor. 2016. "ENmix: A Novel Background Correction Method for Illumina HumanMethylation450 BeadChip." *Nucleic Acids Research* 44 (3): e20.

Yaari, Gur, Christopher R. Bolen, Juilee Thakar, and Steven H. Kleinstein. 2013. "Quantitative Set Analysis for Gene Expression: A Method to Quantify Gene Set Differential Expression Including Gene-Gene Correlations." *Nucleic Acids Research* 41 (18): e170.

Young, Matthew D., Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. 2010. "Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias." *Genome Biology* 11 (2): R14.

Zhang, Yuxia, Jovana Maksimovic, Gaetano Naselli, Junyan Qian, Michael Chopin, Marnie E. Blewitt, Alicia Oshlack, and Leonard C. Harrison. 2013. "Genome-Wide DNA Methylation Analysis Identifies Hypomethylated Genes Regulated by FOXP3 in Human Regulatory T Cells." *Blood* 122 (16): 2823–36.