BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 15 2013, pages 1851–1857
doi:10.1093/bioinformatics/btt311

Genome analysis

Advance Access publication June 3, 2013

Gene-set analysis is severely biased when applied to genome-wide methylation data

Paul Geeleher[1,2], Lori Hartnett[3], Laurance J. Egan[3], Aaron Golden[4], Raja Affendi Raja Ali[3] and Cathal Seoighe[2,*]

[1]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL 60637 USA, [2]Department of Mathematics, Statistics and Applied Mathematics and [3]Department of Pharmacology and Therapeutics, National University of Ireland, Galway, Ireland and [4]Department of Genetics, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: DNA methylation is an epigenetic mark that can stably repress gene expression. Because of its biological and clinical significance, several methods have been developed to compare genome-wide patterns of methylation between groups of samples. The application of gene set analysis to identify relevant groups of genes that are enriched for differentially methylated genes is often a major component of the analysis of these data. This can be used, for example, to identify processes or pathways that are perturbed in disease development. We show that gene-set analysis, as it is typically applied to genome-wide methylation assays, is severely biased as a result of differences in the numbers of CpG sites associated with different classes of genes and gene promoters.

Results: We demonstrate this bias using published data from a study of differential CpG island methylation in lung cancer and a dataset we generated to study methylation changes in patients with long-standing ulcerative colitis. We show that several of the gene sets that seem enriched would also be identified with randomized data. We suggest two existing approaches that can be adapted to correct the bias. Accounting for the bias in the lung cancer and ulcerative colitis datasets provides novel biological insights into the role of methylation in cancer development and chronic inflammation, respectively. Our results have significant implications for many previous genome-wide methylation studies that have drawn conclusions on the basis of such strongly biased analysis.

Contact: cathal.seoighe@nuigalway.ie

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 26, 2013; revised on April 29, 2013; accepted on May 24, 2013

1 INTRODUCTION

Microarrays and high-throughput sequencing are frequently used to assess the methylation status of CpG sites and CpG islands genome-wide. Array platforms for this purpose have been developed by Agilent, Illumina and NimbleGen, and several high-throughput sequencing-based methods have also been developed, such as genome-wide bisulphate sequencing, meDIP-seq (Weber *et al.*, 2005) and HELP-seq (Oda *et al.*, 2009). Gene-set analysis (GSA) is frequently used to discover meaningful biological patterns from lists of genes generated from high-throughput experiments, including genome-wide DNA methylation studies. The objective is typically to identify similarities between the genes, with respect to annotations available from sources such as the Gene Ontology (GO) (Ashburner *et al.*, 2000) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). Popular tools for this purpose include *GOstats* (Falcon and Gentleman, 2007) and *DAVID* (Huang *et al.*, 2009a, b). A significant result for a gene set is interpreted as evidence that the corresponding biological function or process is affected in the experimental condition or treatment. A key assumption of GSA methods is that all genes have, *a priori*, the same probability of appearing in the list. If this is not true, that is, if certain classes of genes are more likely to appear in the list, regardless of the treatments or conditions being investigated, this has the potential to cause misleading results from GSA.

The application of GSA to lists of genes found to be differentially expressed between groups of samples using RNA-seq is known to be affected by such a bias (Oshlack and Wakefield, 2009). This is because there is more power to detect changes in expression for genes with higher numbers of mapped sequence reads. Consequently, longer and more highly expressed genes (which tend to have more mapped reads) are more likely to be identified as differentially expressed. The lengths and expression levels of genes frequently differ between gene sets, resulting in some gene sets wrongly appearing to be enriched among the differentially expressed genes. A method to correct this bias has already been developed (Young *et al.*, 2010). However, the application of GSA is not restricted to the results of high-throughput gene expression measurements; the same approach is used for many other high-throughput experiments. A similar issue has also been highlighted in ChIP-Seq data, when distal binding sites are included in the analysis. In this case, GSA may be confounded when genes in different gene sets are represented by different proportions of the genome. McLean *et al.* (2010) developed GREAT, a tool that can take account of these differences by using a binomial test over a user-defined set of genomic regions. However, GREAT is not suitable for analysis of methylation data, as it cannot differentiate between

*To whom correspondence should be addressed.

one gene in a set with many highly differentially methylated CpG sites or many genes in the same set with one highly differentially methylated CpG site each. This distinction is important in studies of DNA methylation because there are many cases where methylation of only one CpG site has been shown to perturb expression (for example, Claus *et al*., 2012; Deng *et al*., 1999; Sohn *et al*., 2010; Zou *et al*., 2006); hence, GREAT has not been applied to data of this type. Here, we focus on the application of GSA to the results of high-throughput DNA methylation experiments. We show that GSA, as it is typically applied to DNA methylation data, is severely biased and show that methods that have previously been applied to RNA-seq data can be adapted to correct this bias.

2 RESULTS AND DISCUSSION

2.1 Bias in GSA applied to the results of genome-wide differential methylation studies

Microarray platforms designed to profile DNA methylation across the genome are typically designed such that some genes are associated with a large number of probes, whereas others have few associated probes. These differences stem from the fact that different genes and gene promoters contain different numbers of CpG sites. On the Agilent Human CpG Island array, for example, the number of probes per gene promoter ranges from 1 to 285 (Supplementary Fig. S1). Similar platforms by NimbleGen (Human DNA Methylation 385 K Promoter Plus CpG Island Array) and Illumina (Infinium HumanMethylation450 BeadChip) contain from 1 to 80 and 1 to 1288 probes per gene, respectively (Supplementary Fig. S2). The reason that these differences are problematic for GSA becomes clear when we consider the methods currently used to identify differentially methylated genes. For example, comparing tumor and normal lung samples, Helman *et al*. (2012) called a gene differentially methylated when two nearby probes (allowing one intervening probe) both showed at least a 2-fold difference in methylation between the sample groups. Given this criterion, genes (for example, *SOCS4*) that have only two associated probes will be tested just once for differential methylation. By contrast, *PPP2R3B*, which has 197 associated probes, is tested nearly 400 times. It is clear that *PPP2R3B* is far more likely to give false-positive results (resulting from probes exceeding these fixed thresholds by chance). Indeed, there is also more power to detect a real differential methylation signal for genes with many associated probes.

Many other *ad hoc* criteria have been applied to define differentially methylated genes. For example, Dunwell *et al*. (2010) classed genes as differentially methylated if one associated probe reached a fold change of >3. Other authors have used more complex experimental designs and data analysis methodologies, but these do not eliminate the bias. For example, Kalari *et al*. (2012) used a peak calling method to identify regions that appeared enriched for methylation; however, because of the different numbers of probes associated with each gene, it is clearly more likely to call peaks on genes with many associated probes. In all cases, where some genes are tested many more times than others (as is typical in methylation analysis), genes with more associated probes (in the case of microarrays) or more associated CpG sites (in the case of high-throughput sequencing) are more likely to fulfill whatever criteria is used, violating a key assumption of GSA. The reason that this causes such a strong bias in GSA is because there are also large differences between gene sets in the mean numbers of associated probes per gene. For example, on the Agilent Human CpG Island microarray, genes annotated to the Gene Ontology term 'Embryonic organ development' (GO:0048568) have, on average, 22.7 associated probes, more than twice the average of 9.8 probes for all genes. Consequently, if the methodology used to identify differentially methylated promoters is sensitive to the number of probes in the promoter, this gene set is likely to contain a disproportionate number of significant genes, and thus is more likely to appear to be significantly enriched in the subsequent GSA.

2.2 Reanalysis of a published dataset

Many published studies have applied GSA to high-throughput methylation data (for example, Booth *et al*., 2012; Deng *et al*., 2009; Doi *et al*., 2009; Elango *et al*., 2009; Irizarry *et al*., 2009; Liu *et al*., 2010; Schroeder *et al*., 2011; Sen *et al*., 2010; Sproul *et al*., 2012; Takeshima *et al*., 2009; Zhu *et al*., 2012). In all of these studies, regardless of the platform used, gene sets relating to development, transcription or differentiation were reported to be enriched for differentially methylated genes; however, these gene sets are typically associated with large numbers of probes per gene (the precise number depending on platform; Fig. 1a). We now consider one such study in detail. Rauch *et al*. (2008) used the Agilent Human CpG Island microarray to assess methylation in five lung cancer samples compared with normal lung tissue and Helman *et al*. (2012) applied GSA to identify hypermethylated gene sets in this dataset. CpG islands were called as hypermethylated in a sample 'when at least two adjacent probes, allowing a one-probe gap, within the CpG island scored a fold-difference factor of >2 when comparing tumor and normal tissue DNA' (Rauch *et al*., 2008). Genes were considered hypermethylated in lung cancer if any associated CpG island met this criterion in four out of the five samples. The R package *GOstats* was then used to assess aberrant methylation of GO biological processes (BP) containing between 100 and 1000 genes (Rauch *et al*., 2008). We followed the original methodology and obtained results similar to Helman *et al*. (2012) (Table 1). The most significantly enriched gene sets (including differentiation/developmental and transcription factor activity related gene sets) all consisted of genes that were associated with far more microarray probes than average (Fig. 1b). Furthermore, the enrichment *P*-value for each gene set was strongly correlated with the mean number of probes per gene (Fig. 1c). This suggests that these results may be attributable, at least in part, to differences between gene sets in the number of probes per gene.

2.3 Demonstration of bias using randomized probe locations

As an additional line of evidence, we show that significant results can be achieved from this dataset using randomized data. To do this, we carried out 100 random permutations of the probe log-intensity values within each sample and repeated the inference of differential methylation followed by GSA. To ensure that the results were comparable with the original data, we modified

the fold change cut-off for differential methylation so that the average number of hypermethylated genes was the same as in the original data. Altering the fold change cut-off was necessary because the methylation status of adjacent CpG sites is correlated in the real data (Bell *et al*., 2011), resulting in a higher probability of adjacent CpG sites showing differential methylation between groups. If the application of GSA to this dataset was valid, it should not reveal any significant results because in the permuted data, each gene was associated with a random set of probes. Yet, when we applied GSA to the genes found to be differentially methylated in this random dataset, many of the same gene sets that were reported in the original analysis were found to be highly significant (Fig. 1d and Supplementary Table S1) in the vast majority of the random permutations.

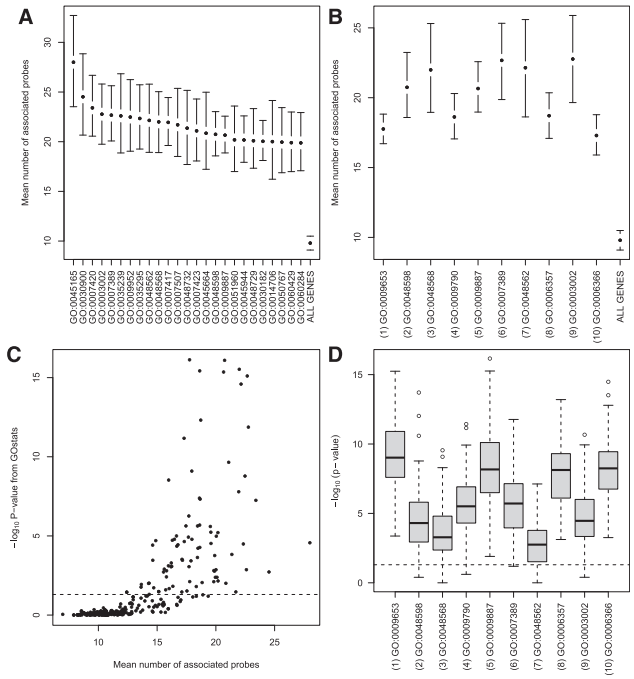

Fig. 1. (**A**) The 25 GO BPs (of size >100 and <1000) with the largest numbers of associated probes per genes on average (on the Agilent Human CpG Island array). All of these 25 gene sets are related to development, differentiation or transcription factor activity. (**B**) The average numbers of probes associated with the top 10 gene sets from the uncorrected GSA on the lung cancer data. In all cases, these gene sets have many more associated probes than average ($P < 2.2 \times 10^{-16}$ in all cases). (**C**) $-\log_{10}$ *P*-values for gene sets in the lung cancer study plotted against average number of associated probes per gene in the gene set. (**D**) Boxplots of $-\log_{10}$ *P*-values for the top 10 GO BPs in the lung cancer dataset obtained from 100 random permutations of probe locations. The dashed line shows the $P = 0.05$ threshold

2.4 Demonstration of bias using sample randomization

An alternative way to test the validity of results obtained using GSA is through permutation of sample labels (Barry *et al*., 2005; Efron and Tibshirani, 2007). If similar GSA results are obtained from random groupings of the samples, then the GSA results do not reflect real biological differences between the sample groups (e.g. tumor versus normal), but rather are an artifact of the analysis. High-throughput methylation experiments typically do not admit straightforward label permutation because different samples are often hybridized to different channels of the same microarray. This is the case for the data of Helman *et al*. (2012) that was generated by hybridizing tumor and normal tissue to dual channel arrays. However, it is possible to reproduce the effect of sample label permutation by inverting the fold change values for each pair of samples. The data consisted of five pairs of samples, resulting in 32 possible configurations when probe intensities are inverted in this way. For each permutation, we ranked genes by methylation fold change of the most differentially methylated probe and applied GSA to the same number of genes as in the original analysis. Again we found that the same gene sets that were reported in the original analysis also seem to be enriched for the inverted datasets (Fig. 2a). In fact 6 of the 10 most significant gene sets were significant for all of the configurations.

2.5 GSA applied to differential methylation in ulcerative colitis

As a further demonstration of the bias, we analyzed data generated again using the Agilent CpG Island microarray, this time applied to sigmoid colon biopsies from five patients with long standing (>25 years) ulcerative colitis (UC), and five healthy age-matched controls (manuscript under review). This is a

Table 1. Top 10 GO BP categories for uncorrected GSA on the lung cancer dataset

GOBPID	Count	Expected count	*P*-value	FDR	Term
GO:0009653	35	7.40	7.48×10^{-17}	1.18×10^{-14}	Anatomical structure morphogenesis
GO:0048598	21	1.94	8.07×10^{-17}	1.18×10^{-14}	Embryonic morphogenesis
GO:0048568	17	1.10	2.98×10^{-16}	2.59×10^{-14}	Embryonic organ development
GO:0009790	25	3.36	3.74×10^{-16}	2.59×10^{-14}	Embryonic development
GO:0009887	25	3.38	4.42×10^{-16}	2.59×10^{-14}	Organ morphogenesis
GO:0007389	19	1.63	7.88×10^{-16}	3.85×10^{-14}	Pattern specification process
GO:0048562	15	0.85	2.56×10^{-15}	1.07×10^{-13}	Embryonic organ morphogenesis
GO:0006357	24	4.15	4.84×10^{-13}	1.77×10^{-11}	Regulation of transcription from RNA polymerase ii promoter
GO:0003002	15	1.29	1.34×10^{-12}	4.36×10^{-11}	Regionalization
GO:0006366	25	5.15	6.79×10^{-12}	1.99×10^{-10}	Transcription from RNA polymerase II promoter

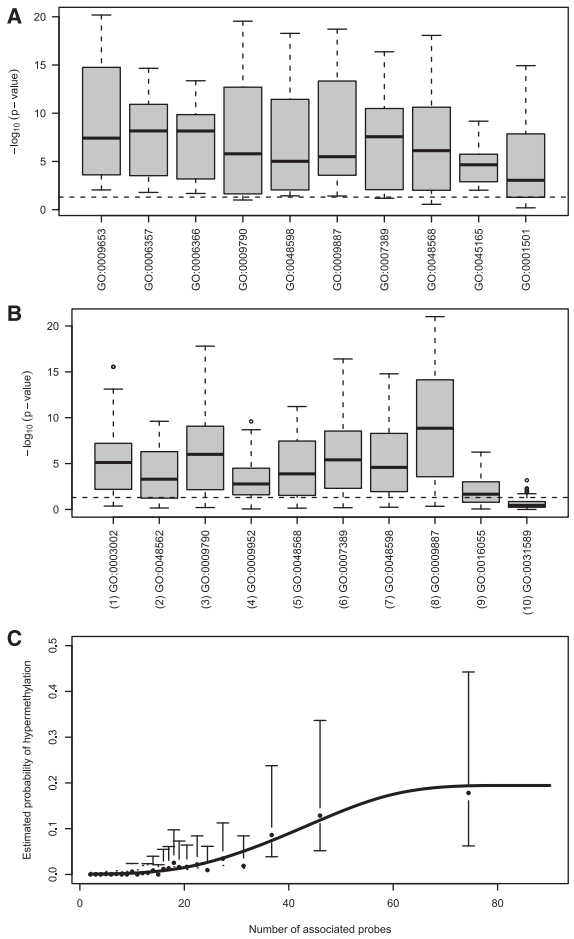

Fig. 2. (**A**) Boxplot of *P*-values from sample label inversions on the lung cancer data. (**B**) Boxplot of *P*-values from all possible sample label permutations on the UC data. (**C**) Fit of GOseq's six knot monotonic spline function to the lung cancer data. The spline is shown as the solid black line. The points show the proportion of hypermethylated genes, in bins of minimum size 100 genes. The 95% confidence intervals for the bins are shown as vertical lines

different experimental design to the dataset discussed earlier in the text, as cases and control samples were hybridized to different microarrays. The Cy3 channels were hybridized with immunoprecipitated methylated DNA [isolated using the MeDIP (Mohn *et al.*, 2009) approach], and the Cy5 channels were hybridized with input DNA (see Section 4). Thus, the log intensity ratio of a probe is, in this case, indicative of the extent of methylation of a probe in a given sample, rather than a log ratio of two different samples (as in the lung cancer experiment). This experimental design allows conventional label permutations. To select a candidate list of differentially methylated genes in this dataset, we used the Bioconductor package *limma* (Smyth, 2005) and identified probes that were hypermethylated in the UC samples ($P<0.05$). All genes associated with at least one hypermethylated probe were considered hypermethylated (see Section 4). This approach identified a foreground list of 380 genes for gene set analysis, which was carried out using *GOstats* (also for GO BPs with between 100 and 1000 associated genes). Once again, the results contain a large number of biological processes related to development and differentiation (Table 2).

Interestingly, the GSA results obtained for this dataset were highly correlated with the results of the lung cancer dataset (Pearson $r=0.69$, $P<2.2\times10^{-16}$). This could suggest that similar biological processes are perturbed in the two cases, which would be of interest, as long-standing UC may be associated with the development of colon cancer (Eaden *et al.*, 2001). However, because of the bias we have identified, this interpretation could be misleading, and the correlation may simply be the result of a shared artifact. Indeed, for the UC dataset we found that a large majority the same gene sets identified earlier in the text remained highly significant when sample labels were randomized, with in many cases *P*-values more extreme than 10^{-20} generated from arbitrary arrangements of the samples (Fig. 2B). This shows that even when using a different experimental design and different criteria to define methylation, highly significant results can again be achieved from random data.

2.6 Bias correction

We applied the Bioconductor package *GOseq* (Young *et al.*, 2010) to control for the fact that different genes have, *a priori*, different probabilities of appearing in the foreground list. This package was developed to control for bias in GSA applied to RNA-seq data. The package uses (by default) a six knot monotonic spline function to model the association between the odds of a gene appearing in the foreground list (i.e. being detected as differentially expressed, or in our case hypermethylated) and the value of the confounding variable (in our case the number of CpG probes associated with the gene; (Fig. 2C). The model can be used to predict the probability of a gene appearing in the foreground list, as a function of the number of associated probes. Given these probabilities, empirical *P*-values are calculated by simulation.

In the lung cancer dataset, the number of GO BP categories that were significantly enriched ($FDR<0.05$) after correction was much smaller than previously (9, compared with 72). After correction (Table 3), 'Embryonic organ morphogenesis' was the most significant category ($P=4.9\times10^{-4}$, compared with $P=8.1\times10^{-17}$, before correction). The expected number of hypermethylated genes for 'Embryonic organ morphogenesis' rose from 1.9 to 3.4, but this is still considerably fewer than the observed number of hypermethylated genes for this category, which is 15. This suggests that the reported hypermethylation of developmental associated genes in this lung cancer dataset is not an artifact of the higher numbers of associated probes. However, several of the gene sets identified in the original analysis are no longer significant. These include gene sets related to transcription factor activity and, perhaps importantly, the gene sets related directly to differentiation. The *P*-value for 'Regulation of cell differentiation' increased from $P=3.0\times10^{-4}$ to $P=0.38$ and 'Cell morphogenesis involved in differentiation' from $P=8.8\times10^{-3}$ to $P=0.48$. This brings into question the validity of the original conclusions of Helman *et al.* that hypermethylation silences genes required for maintenance of the differentiated state. Results for all GO BP terms evaluated are provided as Supplementary Table S2.

Table 2. Top 10 GO BP categories for uncorrected GSA on the UC dataset

GOBPID	Count	Expected count	*P*-value	Term
GO:0003002	14	4.23	8.52×10^{-05}	Regionalization
GO:0048562	11	2.79	1.04×10^{-04}	Embryonic organ morphogenesis
GO:0009790	25	11.30	1.57×10^{-04}	Embryonic development
GO:0009952	11	3.06	2.37×10^{-04}	Anterior/posterior pattern formation
GO:0048568	11	3.59	9.44×10^{-04}	Embryonic organ development
GO:0007389	14	5.43	1.10×10^{-03}	Pattern specification process
GO:0048598	15	6.38	1.86×10^{-03}	Embryonic morphogenesis
GO:0009887	21	11.03	3.45×10^{-03}	Organ morphogenesis
GO:0016055	10	3.67	3.73×10^{-03}	Wnt receptor signaling pathway
GO:0031589	8	2.54	3.90×10^{-03}	Cell-substrate adhesion

Table 3. GO BP categories with FDR < 0.05 from a GSA corrected using *GOseq* on the lung cancer dataset

GOBPID	Count	Expected count	*P*-value	FDR	Term
GO:0048562	15	3.38	1.67×10^{-06}	4.90×10^{-04}	Embryonic organ morphogenesis
GO:0048568	17	4.14	4.21×10^{-06}	6.16×10^{-04}	Embryonic organ development
GO:0003002	15	4.49	4.40×10^{-05}	3.63×10^{-03}	Regionalization
GO:0009887	25	10.38	6.00×10^{-05}	3.63×10^{-03}	Organ morphogenesis
GO:0048598	21	6.27	6.20×10^{-05}	3.63×10^{-03}	Embryonic morphogenesis
GO:0009790	25	11.01	1.18×10^{-04}	5.40×10^{-03}	Embryonic development
GO:0007423	13	3.86	1.29×10^{-04}	5.40×10^{-03}	Sensory organ development
GO:0007389	19	5.31	2.81×10^{-04}	1.03×10^{-02}	Pattern specification process
GO:0009952	11	3.39	5.06×10^{-04}	1.65×10^{-02}	Anterior/posterior pattern formation

2.7 Validation of bias-corrected GSA by comparison with label permutation

Several authors have previously suggested correcting the results of GSA using sample label permutations (Barry *et al.*, 2005; Efron and Tibshirani, 2007). For example, the popular tool GSEA (Subramanian *et al.*, 2005) uses this approach. In many cases (for example, the lung cancer study discussed earlier in the text), this is not possible, and in many other cases, it may be highly computationally intensive, there may be a limited number of samples (meaning that it will be impossible to achieve statistical significance), or it may be difficult because of the complexity of the analysis pipelines sometimes applied to methylation datasets. However, in the case of our UC dataset, it is straightforward. Thus, we used this dataset to compare the results of the *GOseq*-corrected GSA method to the results obtained from sample label permutations.

Applying an uncorrected GSA to the UC dataset identifies many highly significant gene sets (Table 2). After correction using sample label permutations, 19 gene sets were identified with $P < 0.05$ (Table 4). GSA corrected by *GOseq* identified 12 gene sets (Table 5; Supplementary Fig. S3) compared with 262 when using an uncorrected approach (suggesting that a large proportion of the results from the original analysis were artifacts). Of the 12 gene sets identified using the corrected GSA, 11 were also identified by label permutation. The similarity to the results of a robust permutation-based approach provides good evidence that the corrected GSA performs well and is a suitable method of accounting for bias in methylation data. By comparison, only 12 of the 19 gene sets identified using label permutation were detected using the uncorrected GSA, despite an order of magnitude more processes reported significant in the latter analysis. When a corrected analysis is applied, results in the UC dataset are not significant after correction for multiple testing. Where label permutation is possible (e.g. in the UC experimental design), it can be used to perform GSA in a way that is robust to the differences in the number of probes per gene. However, it is likely that *GOseq* provides better power to detect gene sets that are enriched for differentially methylated genes, as the statistical significance that can be achieved by a permutation method can be limited because of the relatively small sample sizes that are often encountered in genome-wide methylation experiments.

In the previous section, we showed that the results of an uncorrected GSA applied to the lung cancer and UC datasets were similar, a result that could potentially provide insight into UC-associated carcinogenesis. However, when we corrected for the number of probes per gene, the Pearson correlation between the *P*-values in the UC and lung cancer datasets is dramatically lower than for the uncorrected results ($r = 0.17$, compared with $r = 0.69$), suggesting that the similar results were largely artifactual.

3 CONCLUSIONS

In general, when different genes and gene sets are associated with different *a priori* probabilities of appearing in the foreground list as a consequence of factors other than those that are of biological interest there is the potential for bias. This arises in many GSA applications. It is common in GSA to associate multiple and different numbers of features with each gene; typically, multiple features are collapsed onto single-gene identifiers. For instance, the popular web-based GSA tool, *DAVID* (Huang *et al*., 2009a, b) offers the option to use microarray probe IDs (e.g. from methylation or gene expression arrays) as foreground and background lists. These are converted to unique gene IDs before statistical analysis. When there is a difference in the number of probes associated with each gene, this can give rise to the bias that we have outlined. In this article, we have demonstrated that this causes severely biased results when GSA is applied to high-throughput methylation data, typically leading to false-positive results for gene sets related to development, differentiation and transcription. This bias can be corrected by applying a GSA method, such as *GOseq*, that models the relationship between the number of features (e.g. CpG probes in the case of microarrays or CpG sites in the case of high-throughput sequencing) associated with a gene and its probability of appearing in the foreground list, or where applicable, by using sample label permutations.

Table 4. Top 10 GO BPs from a GSA corrected using label permutation (UC dataset)

GOBPID	Count	*P*-value	Term
GO:0009952	11	7.94×10^{-03}	Anterior/posterior pattern formation
GO:0008610	12	7.94×10^{-03}	Lipid biosynthetic process
GO:0006629	20	7.94×10^{-03}	Lipid metabolic process
GO:0031589	8	1.59×10^{-02}	Cell-substrate adhesion
GO:0043085	17	2.38×10^{-02}	Positive regulation of catalytic activity
GO:0044255	14	2.38×10^{-02}	Cellular lipid metabolic process
GO:0003002	14	3.17×10^{-02}	Regionalization
GO:0016055	10	3.17×10^{-02}	Wnt receptor signaling pathway
GO:0006644	7	3.17×10^{-02}	Phospholipid metabolic process
GO:0019637	7	3.17×10^{-02}	Organophosphate metabolic process
GO:0030003	8	3.17×10^{-02}	Cellular cation homeostasis

4 METHODS

4.1 Gene-set analysis using label permutation

Sample labels were rearranged in all possible combinations. As there were a total of 10 samples, split equally between UC and control phenotypes, this yielded 126 distinct arrangements of the samples. For each of these arrangements, the enrichment odds ratio test statistic was re-calculated for each gene set, using the same pipeline as the original analysis and selecting the same number of hypermethylated genes (so that the GSA is comparable between the observed and permuted data). *P*-values were calculated as the proportion of the test statistics that were as extreme, or more extreme, than the test statistic corresponding to the observed data.

4.2 UC microarray data

MeDIP was performed to capture methylated DNA sequence as previously described by Weber *et al*. with slight modifications. Briefly, 10 µg of 5-methylcytosine antibody was incubated with 50 µl of Dynabeads M-28 Sheep anti-mouse IgG for 5 h in immunoprecipitate (IP) buffer at 4°C. Genomic DNA was sonicated using the Branson digital sonifier, and 4 µg of genomic DNA was incubated with the antibody-beads complex overnight at 4°C. Then, the DNA-antibody-dynabeads complex was washed three times with IP buffer and incubated with 5 µl of proteinase K for 2 h at 55°C. In our experiment, we labeled the IP DNA with fluorescent dye, cyanine 3 and reference (R) DNA with cyanine 5 and co-hybridized to the Agilent microarrays. The MeDIP followed by CpG island microarray analysis enables us to identify the methylated and unmethylated CpG islands between long-standing UC patients and age-matched control patients. Purification of labeled products, array hybridization and scanning were performed at the functional genomics and high-throughput screening facility at the National Centre for Biomedical Engineering Science, NUI Galway.

Data were quantile normalized and analyzed using the Bioconductor library *limma*. Genes with at least one associated hypermethylated probe in UC ($P < 0.05$) were selected for GSA. These data have been uploaded to GEO and are available under accession number *GSE39188*.

Table 5. Top 10 GO BPs from a GSA corrected using *GOseq* on the UC dataset

GOBPID	Count	Expected count	*P*-value	Term
GO:0008610	12	5.39	7.70×10^{-03}	Lipid biosynthetic process
GO:0031589	8	3.27	1.38×10^{-02}	Cell-substrate adhesion
GO:0006644	7	2.82	2.41×10^{-02}	Phospholipid metabolic process
GO:0019637	7	2.96	2.95×10^{-02}	Organophosphate metabolic process
GO:0043085	16	9.88	3.59×10^{-02}	Positive regulation of catalytic activity
GO:0048562	11	6.07	3.70×10^{-02}	Embryonic organ morphogenesis
GO:0044087	7	3.28	4.00×10^{-02}	Regulation of cellular component biogenesis
GO:0016053	6	2.56	4.02×10^{-02}	Organic acid biosynthetic process
GO:0046394	6	2.56	4.02×10^{-02}	Carboxylic acid biosynthetic process
GO:0006629	20	13.12	4.29×10^{-02}	Lipid metabolic process

ACKNOWLEDGEMENT

The authors thank Dr John Newell for advice on statistical analysis.

Funding: P.G. was supported by the Irish Research Council for Science, Engineering and Technology (IRCSET) Embark Initiative's Enterprise Partnership Scheme. C.S. is supported by Science Foundation Ireland (07/SK/M1211b).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*., **25**, 25–29.

Barry,W. *et al*. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Bell,J. *et al*. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol*., **12**, R10.

Booth,M.J. *et al*. (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, **336**, 934–937.

Claus,R. *et al*. (2012) Quantitative DNA methylation analysis identifies a single CpG dinucleotide important for ZAP-70 expression and predictive of prognosis in chronic lymphocytic leukemia. *J. Clin. Oncol*., **30**, 2483–2491.

Deng,G. *et al*. (1999) Methylation of CpG in a small region of the hMLH1 promoter invariably correlates with the absence of gene expression. *Cancer Res*., **59**, 2029–2033.

Deng,J. *et al*. (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat. Biotechnol*., **27**, 353–360.

Doi,A. *et al*. (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet*., **41**, 1350–1353.

Dunwell,T. *et al*. (2010) A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol. Cancer*, **9**, 44.

Eaden,J. *et al*. (2001) The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut*, **48**, 526–535.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat*., **1**, 107–129.

Elango,N. *et al*. (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proc. Natl Acad. Sci. USA*, **106**, 11206–11211.

Falcon,S. and Gentleman,R. (2007) Using gostats to test gene lists for go term association. *Bioinformatics*, **23**, 257–258.

Helman,E. *et al*. (2012) DNA hypermethylation in lung cancer is targeted at differentiation-associated genes. *Oncogene*, **31**, 1181–1188.

Huang,D.W. *et al*. (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*., **37**, 1–13.

Huang,D.W. *et al*. (2009b) Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc*., **4**, 44–57.

Irizarry,R.A. *et al*. (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet*., **41**, 178–186.

Kalari,S. *et al*. (2012) The DNA methylation landscape of small cell lung cancer suggests a differentiation defect of neuroendocrine cells. *Oncogene* [Epub ahead of print, doi: 10.1038/onc.2012.362, August 20, 2012].

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.

Liu,J. *et al*. (2010) A study of the influence of sex on genome wide methylation. *PLoS One*, **5**, e10028.

McLean,C.Y. *et al*. (2010) Great improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol*., **28**, 495–501.

Mohn,F. *et al*. (2009) Methylated DNA immunoprecipitation (medip). *Methods Mol. Biol*., **507**, 55–64.

Oda,M. *et al*. (2009) High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res*., **37**, 3829–3839.

Oshlack,A. and Wakefield,M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*, **4**, 14.

Rauch,T.A. *et al*. (2008) High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. *Proc. Natl Acad. Sci. USA*, **105**, 252–257.

Schroeder,D.I. *et al*. (2011) Large-scale methylation domains mark a functional subset of neuronally expressed genes. *Genome Res*., **21**, 1583–1591.

Sen,G.L. *et al*. (2010) DNMT1 maintains progenitor function in self-renewing somatic tissue. *Nature*, **463**, 563–567.

Smyth,G.K. (2005) *Limma: Linear Models for Microarray Data*. Springer, New York, pp. 397–420.

Sohn,B.H. *et al*. (2010) Functional switching of TGF-beta1 signaling in liver cancer via epigenetic modulation of a single CpG site in TTP promoter. *Gastroenterology*, **138**, 1898–1908.

Sproul,D. *et al*. (2012) Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biol*., **13**, R84.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Takeshima,H. *et al*. (2009) The presence of RNA polymerase II, active or stalled, predicts epigenetic fate of promoter CpG islands. *Genome Res*., **19**, 1974–1982.

Weber,M. *et al*. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet*., **37**, 853–862.

Young,M.D. *et al*. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*., **11**, R14.

Zhu,J.G. *et al*. (2012) Differential DNA methylation status between human preadipocytes and mature adipocytes. *Cell Biochem. Biophys*., **63**, 1–15.

Zou,B. *et al*. (2006) Correlation between the single-site CpG methylation and expression silencing of the XAF1 gene in human gastric and colon cancers. *Gastroenterology*, **131**, 1835–1843.