Clinical Epigenetics

METHODOLOGY Open Access

An empirically driven data reduction method on the human 450K methylation array to remove tissue specific non-variable CpGs

Rachel D. Edgar, Meaghan J. Jones, Wendy P. Robinson and Michael S. Kobor*

Abstract

Background: Population based epigenetic association studies of disease and exposures are becoming more common with the availability of economical genome-wide technologies for interrogation of the methylome, such as the Illumina 450K Human Methylation Array (450K). Often, the expected small number of differentially methylated cytosine-guanine pairs (CpGs) in studies of the human methylome presents a statistical challenge, as the large number of CpGs measured on the 450K necessitates careful multiple test correction. While the 450K is a highly useful tool for population epigenetic studies, many of the CpGs tested are not variable and thus of limited information content in the context of the study and tissue. CpGs with observed lack of variability in the tissue under study could be removed to reduce the data dimensionality, limit the severity of multiple test correction and allow for improved detection of differential DNA methylation.

Methods: Here, we performed a meta-analysis of 450K data from three commonly studied human tissues, namely blood (605 samples), buccal epithelial cells (121 samples) and placenta (157 samples). We developed lists of CpGs that are non-variable in each tissue.

Results: These lists are surprisingly large (blood 114,204 CpGs, buccal epithelial cells 120,009 CpGs and placenta 101,367 CpGs) and thus will be valuable filters for epigenetic association studies, considerably reducing the dimensionality of the 450K and subsequently the multiple testing correction severity.

Conclusions: We propose this empirically derived method for data reduction to allow for more power in detecting differential DNA methylation associated with exposures in studies on the human methylome.

Keywords: Non-variable, 450K, Tissue, Filter, Power, Multiple-test correction, DNA methylation, Dimensionality reduction

Background

Population studies that interrogate epigenetic signatures associated with environmental variation and disease are becoming increasingly common. The challenge with the majority of epigenome wide association studies (EWAS) of environment and disease is that the epigenetic signals, in terms of detectable number of epigenetic changes and the effect size of changes, between groups are relatively small compared to those observed in EWAS of development, tissues or cancer. Therefore careful and specific methodological steps need to be implemented in analyses to separate any true biological signal from stochastic variation in DNA methylation (DNAm), a phenomenon commonly referred to as noise [1].

One of the most common types of population based epigenetic studies is the examination of DNAm using the Illumina Infinium 450K array (450K) or its related arrays [2]. The Illumina series of DNAm arrays, while highly useful as tools for epigenetic studies, were not designed for any specific human tissue, and a large number of cytosine-guanine pairs (CpGs) lack variability within single tissue studies on the arrays [3–8]. CpGs that are non-variable in a study of a specific disease or tissue

* Correspondence: msk@cmmt.ubc.ca
Department of Medical Genetics, BC Children's Hospital, University of British Columbia, Vancouver, Canada

may be variable in another context and therefore are still valuable on the 450K. However, these tissue specific non-variable CpGs contribute to the high dimensionality of the 450k data and partially necessitate the need for severe multiple test correction. In an effort to rigorously determine the epigenetic signals of environmental exposure and/or disease phenotypes, dimensionality reduction techniques are often employed. These include mixture modelling, principal component analysis, weighted gene co-expression network analysis and elastic net models, among others [9–12]. While these techniques are effective for high-dimensional data reduction, they do not take into account the wealth of independent DNAm data available to build empirical data reduction filters. A common data-driven dimensionality reduction technique is to remove non-variable CpGs from within a specific study and then test only variable sites for association with the exposure of interest [3–8]. While this practice can reduce severe multiple test correction penalties, it can introduce a bias toward significant results [13]. A promising alternative from gene expression analyses is to use a filter based on prior biological knowledge from independent data, which can be highly effective in improving sensitivity while maintaining specificity [13].

Here, we have developed an empirically derived data reduction method in the form of CpG lists which are non-variable in independent cohorts of samples from three commonly used human tissues: blood, buccal epithelial cells and placenta. We anticipate these independently identified non-variable CpG lists will be useful for confirmation of a lack of variability at CpGs in 450K studies of interest. As such, our non-variable CpGs might serve as a benchmark to cross-reference CpGs also seen as non-variable in a study of interest so that these CpGs can be filtered prior to differential DNAm analysis. Removal of these independently verified non-variable CpGs should then allow for a reduced multiple testing space and allow for more power to detect differential DNAm in the study of interest. While this approach will be immediately useful for studies of 450K data, it will also provide a blueprint for similar approaches with emerging technologies such as the Illumina EPIC array. Our filtering approach for data reduction is focused on CpG-by-CpG EWAS analyses, which are very common approaches in DNAm analysis. However, this filtering approach also has the potential to improve the performance of other analyses where a strong signal is expected at a small subset of CpGs and noise in the data is a concern. In the context of the rapidly increasing number of DNAm datasets being produced, we have made our code available so that independent non-variable CpG lists can be rapidly developed for other tissues of interest on the 450K and the EPIC as data becomes available.

Methods

Data collection

The tissue datasets were collected from Gene Expression Omnibus (GEO) [14]. In all tissues, cancer samples were excluded, as cancer is associated with high DNAm variability [15]. For individual tissues, there were a range of exclusion terms by which samples were filtered (Additional file 1: Table S2). Exclusion terms were based on whether the term indicated cancerous tissue, a tissue other than the tissue of interest or a species other than human. In general, data was downloaded as non-normalized betas, but in some cases, M values were converted to beta, and normalized data was used. Each tissue dataset was then filtered down to the minimal number of CpGs with DNAm values across all samples of a tissue (blood 469,961 CpGs, buccal epithelial cells 420,374 CpGs and placenta 484,621 CpGs).

Quality control

To remove CpGs and samples that consistently did not perform well on the 450K, CpGs were filtered if greater than 5% of samples had fewer than three beads contributing to the signal across all samples from a tissue. Samples were removed if 2.5% of CpGs in a sample had fewer than three beads contributing to the signal. Samples were also removed if they had low sample-sample correlation compared to all other samples of a tissue. One sample was filtered from placenta and four entire studies were filtered from blood (total of 158 samples from blood; see Fig. 1; Additional file 1: Table S2). The final studies and samples included are listed in Additional file 1: Table S3.

Non-variable calling

To designate a CpG as non-variable in a tissue, a threshold of 5% range in beta values (DNAm level ranging from 0 to 1) between the 10th and 90th percentile was used [16]. While effect sizes as small as 1% are used in EWAS [8, 17, 18], we used a slightly more stringent definition of change in beta of 5% as we are asking only that the population as a whole varies by at least 5% and are not testing an effect size between groups. CpGs with less than 5% reference range of beta values in a single tissue population were considered non-variable in that tissue.

Genomic enrichment

To explore the genomic context of non-variable CpGs, all CpGs were associated with gene features using the annotation described previously [19] and with CpG island features as provided in the Illumina annotation [2]. The count of non-variable CpGs located in each gene feature (promoter, intragenic, 3 prime region and intergenic) and CpG island feature (island, north and south shore, north and south shelf, and no island association)

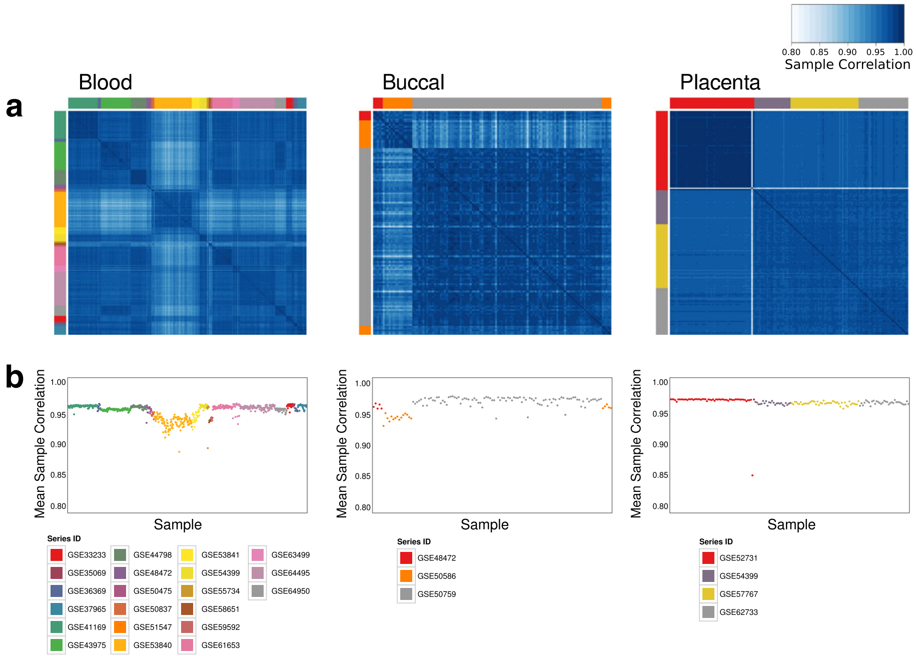

Fig. 1 Quality control of samples from GEO for each tissue type. **a** Heat maps showing sample-sample correlation values. *Side colours* show the study ID of each sample, and samples are ordered by study ID. **b** Plots of the average sample-sample correlation for each sample to show possible outliers and studies with overall low average sample-sample correlation

were compared to the background counts of all CpGs measured, in each tissue. To compare the non-variable CpG counts to the background in each region, 1000 permutations of random CpG lists were used to calculate fold change values over the background [20].

Application of data reduction method

To reproduce the published findings of AHRR DNA methylation changes associated with smoke exposure, a linear modelling approach was used on previously published data [21]. In short, DNAm values were normalized using BMIQ [22], and cell composition was normalized between blood samples [23, 24]. A linear model was run at all CpG sites and delta beta effect sizes were calculated between smokers and non-smokers in the full dataset of 111 blood samples. To simulate a study with reduced power, ten permutations of 24 random samples (12 smokers and 12 non-smokers) were selected and the same linear model was run at all CpGs. To test the data reduction method, the CpGs in the ten smaller cohorts were filtered to 374,945 variable CpGs by overlapping the CpGs that were non-variable in GSE53045 (264,578 CpGs non-variable at a reference range of 0.05) and the blood non-variable CpGs identified in the independent samples (114,204 CpGs described above). Then, the same linear model was run on only variable CpGs. CpGs were associated to genes as previously described [19].

Results

Tissues showed similar levels of non-variable CpGs

DNAm data from publicly available studies was collected for blood, buccal epithelial cells and placenta (21, 3 and 4 studies, respectively). Meta-analysis of samples for each of the tissues showed generally high correlations (70% of sample pairs correlated above 0.95). While there were some samples with higher within study correlations than across study correlations, the overall high correlation of cross study samples can be taken as evidence of the consistency of the 450K across research groups (Fig. 1). While four studies of blood were removed due to low correlation, no obvious explanation of the lack of correlation could be found in the available study characteristic information (Additional file 1: Table S2). The generally high concordance of the DNAm samples from the same tissue but different studies gives us confidence going forward in the appropriateness of comparing variability across studies. After quality control of the data, 605, 121 and 157 samples were used from blood, buccal epithelial cells and placenta, respectively.

A substantial number of tissue-specific non-variable CpGs were identified, thus providing a solid baseline for potential removal from studies of interest to reduce dimensionality. The total number of non-variable CpGs was similar across tissues: blood 114,204 (24%), buccal epithelial cells 120,009 (29%) and placenta 101,367 (21%) and showed a significant overlap of 42,315 non-variable

CpGs (permutation $p < 0.0001$; Fig. 2a). Non-variable CpGs existed in either fully methylated or unmethylated state, with few non-variable CpGs observed at an intermediate DNAm level. In all tissues, the 99th percentile of non-variable CpGs had a mean DNAm greater than 0.80 or less than 0.16 (Fig. 2b). To test robustness of the non-variable CpG lists, we compared the list of non-variable CpGs prior to processing with a similar list generated after normalization or after cell type correction. We found that non-variable CpG lists overlapped by 90% with all processing strategies (Additional file 1).

While exploring the biological role of non-variable CpGs that was not the primary focus of this analysis, we did observe that non-variable CpGs from each tissue were significantly enriched in promoters and CpG islands (relative enrichment = 2.46–8.20, false discovery rate (FDR) = 0.01; Fig. 3), with maximum enrichment in blood and lowest enrichment in placenta. Based on the large overlap in and similar genomic localization of non-variable CpGs between the three tissues, it is likely that the non-variable CpGs identified have similar underlying properties in each tissue.

Application of data reduction method to smoking cohort

To test the utility of our filtering non-variable CpGs as a dimensionality reduction method, capable of improving statistical power and sensitivity, we attempted to demonstrate the gain in statistical power in reproducing a well-accepted true positive DNAm modification associated with smoking. In particular, one of the most reproducible biomarkers in DNAm association studies to date is decreased DNAm associated with smoke exposure at two CpGs in the gene body of AHRR [21, 25–28]. To validate our data reduction method, we used the AHRR signal in response to smoke exposure as a true positive. By reanalyzing all 111 blood samples available with smoking status in the original unfiltered data set (GSE53045) [21], we reproduced the finding of significantly decreased DNAm at two CpGs (cg05575921, cg23576855; FDR <0.05, delta beta 0.1) in AHRR. Interestingly, the non-variable CpGs often reached statistical significance (Fig. 4a), supporting that targeted removal of non-variable CpGs from EWAS improves specificity and reduces spurious associations.

To simulate a less powered study of smoke exposure, we randomly sampled the cohort down to 24 samples (12 smokers, 12 non-smokers) ten times. The same linear model, as used in the full cohort, was run on each of the ten randomly sampled smaller cohorts, but with either all 485,512 CpGs included in the EWAS or with filtering of 110,567 non-variable CpGs (filtered EWAS). This resulted in several interesting insights. First, in nine of the ten low powered EWAS sub samples, the multiple test corrected p values of the two true positive AHRR CpGs of interest were smaller in the filtered data set (Fig. 4b). Second, beyond AHRR, only six out of ten sub samples had any significantly differentially DNAm CpGs regardless of whether we used filtered or unfiltered data (FDR <0.05, delta beta 0.1). Third, in five of these six,

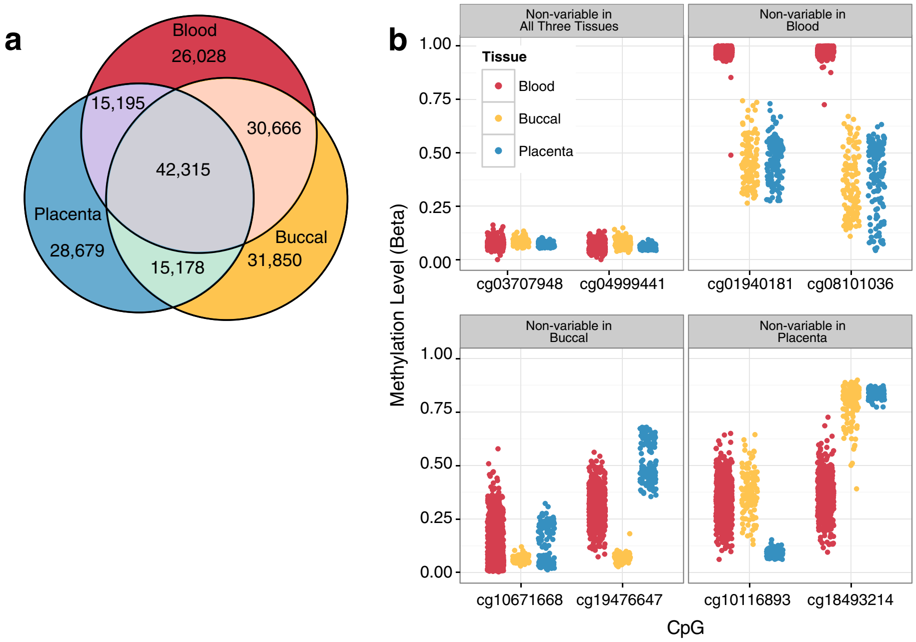

Fig. 2 Non-variable CpGs had similar characteristics in all tissues. **a** Venn diagram showing the overlap of non-variable CpGs between tissues. **b** Methylation levels of representative non-variable CpGs from each tissue

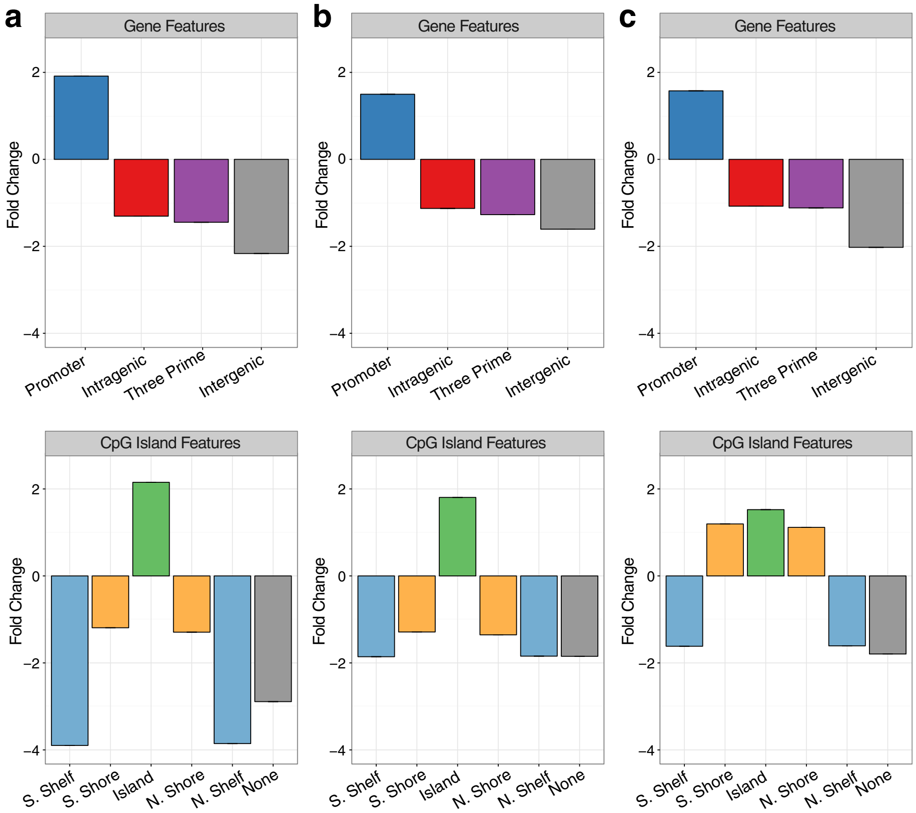

Fig. 3 Non-variable CpGs were enriched in CpG island and promoters. All plots show the enrichment fold change of non-variable CpGs compared to all CpGs available for a tissue. Each pair of plots shows the fold changes in gene regions (*top*) and CpG resort features (*bottom*). **a** Blood non-variable CpGs. **b** Buccal epithelial cell non-variable CpGs. **c** Placenta non-variable CpGs

the filtered data set EWAS resulted in more CpGs with significant differential DNAm. The greater significance of AHRR in the filtered EWAS suggested that filtration of non-variable CpGs should allow for prioritization of true positives, potentially even when the differential DNAm signal is not as strong as AHRR in smoking.

Discussion

Here, we have developed an empirically derived dimensionality reduction method for EWAS, which can reduce noise in 450K data from tissue specific non-variable CpGs. Our proposed method for removing our empirically identified non-variable CpGs is to first confirm if they are also non-variable in the new dataset of interest and remove only those CpGs which are confirmed as non-variable, as presented in the analysis of the AHRR signal in response to smoke exposure. This procedure would avoid removing CpGs that were non-variable in the data collected previously, but do in fact vary in new data being analyzed from the tissue. Generally, previous analyses on 450K data have either filtered based on variability within the study data or not filtered the data on variability at all. We consider our filtration method to be a more moderate compromise between false positive and negatives. Our method is less biased toward false positives than filtering based on variability just in the study data, and also less likely to result in false negatives due to severe multiple test correction when no variability filter is to be used at all [13].

In defining our non-variable CpG list, we were agnostic to normalization methods and did not correct for batch effects between laboratories, beyond removing samples with low sample-sample correlations. We have therefore left in variability in the data due to technical factors that would have been minimized had we combined the data for normalization and performed batch correction. Our list of non-variable CpGs is thus conservative, but should be robust to study specific technical variability, increasing its utility in the community.

We have demonstrated the utility of the filtration in the analysis of smoke exposure in GSE53045, as the successful identification of differential DNAm at the true positive AHRR and the identification of more CpGs genome wide with significantly differentially DNAm. We do not propose simply observing more CpGs with differential DNAm as a good metric for the utility of our data

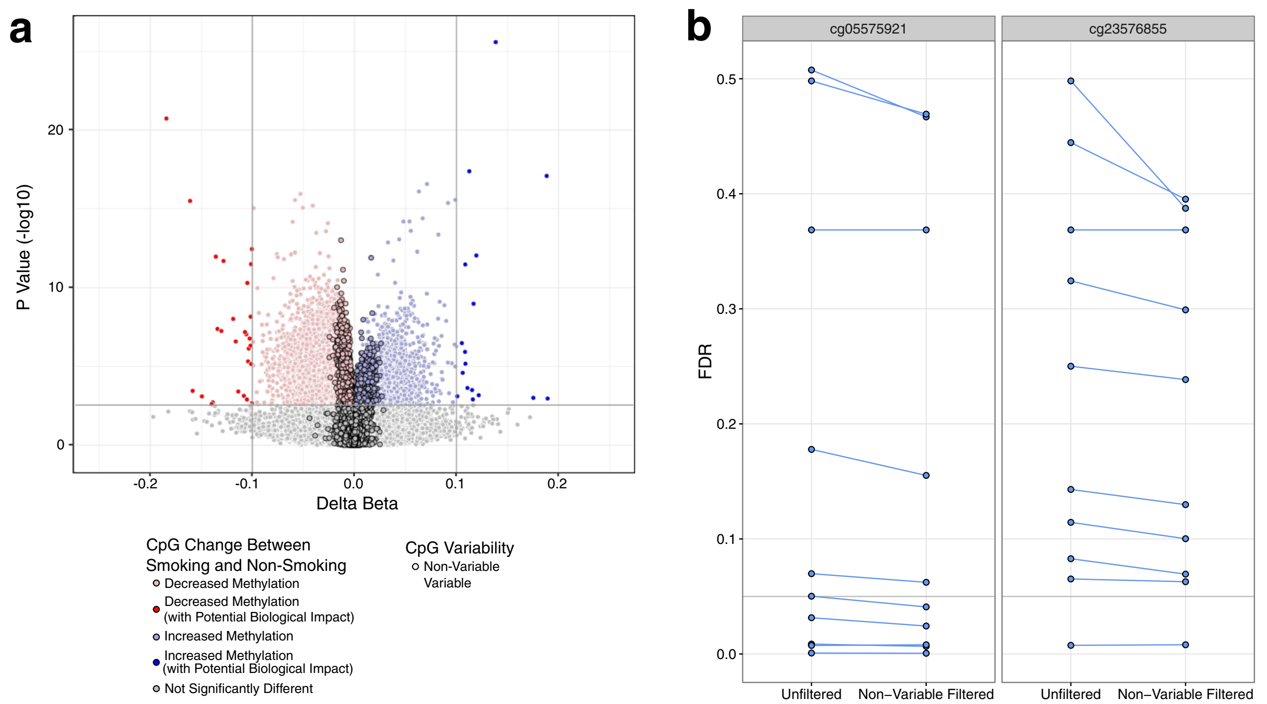

Fig. 4 Multiple test corrected *p* values were lower in the filtered EWAS. **a** Volcano plot of the differential methylation analysis between smoking and non-smoking samples, with no filtering of non-variable CpGs. *Vertical lines* indicate a DNAm difference between 0.1. The *horizontal line* represents an FDR corrected *p* value of 0.05. Points are coloured to highlight CpGs exceeding both the biological and statistical cutoffs. *Points with a black outline* are CpGs found to be non-variable in blood. **b** Shown are the multiple test corrected *p* values (FDR) for the two CpGs of interest in AHRR. Lines connect FDR values between paired permutation sub-samples to show the trend between paired cohorts. The *horizontal line* shows the FDR values of 0.05

reduction method, as some of the significant CpGs identified with filtration will be false positives. However, in combination with the observation of significant differential DNAm at the true positive, AHRR, more consistently with filtration, we are confident that our data reduction method will have utility in allowing identification of replicable differential DNAm in other datasets. Filtering for data reduction will be particularly useful when there is an expectation of CpGs with strong differential methylation signals (>5%); so, the expected magnitude of DNAm change should be carefully considered by the researcher before applying any data reduction. In concert with a stringent biological filter for the change in DNAm level between groups (5–10%) [1, 29], and validation of the 450K results with another technology such as pyrosequencing [1], this tissue specific DNAm data dimensionality reduction method may allow for better and more stringent identification of epigenetic signatures of exposure or disease.

Conclusions

While the ability to define a tissue specific non-variable list will ultimately depend on the amount of data available for the tissue in public repositories, we expect there are already other tissues of interest with sufficient 450K data for which a useful list of non-variable CpGs could be developed. We have therefore made our code for building tissue specific non-variable lists available on GitHub (github.com/redgar598/Tissue_Nonvariable_450K_CpGs). We hope our analysis can be reapplied in the future to update the non-variable CpGs lists for blood, buccal epithelial cells and placenta as more samples become available, and be expanded to more tissues. Additionally, with the increased dimensionality of the newly released Illumina Infinium EPIC array, the need for tissue specific dimensionality reduction will be even greater. The analysis we have outlined and made available can easily be applied to EPIC array datasets as more are released [30].

Additional file

Additional file 1: Table S1. Gene expression omnibus data description and additional analysis. Terms used to exclude samples not of interest in a given tissue. **Table S2.** Quality control filters for each tissue and the resulting final study, sample and CpG numbers. **Table S3.** Series IDs of the final samples used in the meta-analysis of tissue non-variable CpGs. Additional analysis on the stability of the non-variable CpG list with different data processing approaches. (PDF 63 kb)

Abbreviations

450K: Illumina 450K human methylation array; CpG: Cytosine-guanine pair; DNAm: DNA methylation; EWAS: Epigenome wide association studies; FDR: False discovery rate; GEO: Gene expression omnibus

Acknowledgements

The authors would like to thank Dr. Magda Price and Dr. Sara Mostafavi for their thoughts and careful consideration of the approach taken in this work. Thank you to the researchers who kindly deposited their data in public repositories, without their contribution this work would not be possible.

Funding

Support for this work was provided by Canadian Institutes of Health Research (FRN146331); R. Howard Webster Foundation (F13-00031) and W. Garfield Weston Foundation/Brain Canada Foundation (F13-02369).

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Authors' contributions

RDE performed the data analysis and drafted the manuscript. RDE, MJJ, WPR and MSK wrote, reviewed and edited the manuscript. WPR and MSK conceived and coordinated the study. All authors have approved the final version.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 7 October 2016 Accepted: 31 January 2017
Published online: 02 February 2017

References

1. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, Houseman EA, Izzi B, Kelsey KT, Meissner A, Milosavljevic A, Siegmund KD, Bock C, Irizarry RA. Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods. 2013;10(10):949–55.
2. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95.
3. Byun HM, Siegmund KD, Pan F, Weisenberger DJ, Kanel G, Laird PW, Yang AS. Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. Hum Mol Genet. 2009;18(24):4808–17.
4. Glossop JR, Nixon NB, Emes RD, Haworth KE, Packham JC, Dawes PT, Fryer AA, Mattey DL, Farrell WE. Epigenome-wide profiling identifies significant differences in DNA methylation between matched-pairs of T- and B-lymphocytes from healthy individuals. Epigenetics. 2013;8(11): 1188–97.
5. Duong CV, Emes RD, Wessely F, Yacqub-Usman K, Clayton RN, Farrell WE. Quantitative, genome-wide analysis of the DNA methylome in sporadic pituitary adenomas. Endocr Relat Cancer. 2012;19(6):805–16.
6. Fryer AA, Emes RD, Ismail KM, Haworth KE, Mein C, Carroll WD, Farrell WE. Quantitative, high-resolution epigenetic profiling of CpG loci identifies associations with cord blood plasma homocysteine and birth weight in humans. Epigenetics. 2011;6(1):86–94.
7. Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, Kobor MS. Factors underlying variable DNA methylation in a human community cohort. Proc Natl Acad Sci. 2012;109(suppl2):17253–60.
8. Esposito EA, Jones MJ, Doom JR, MacIsaac JL, Gunnar MR, Kobor MS. Differential DNA methylation in peripheral blood mononuclear cells in adolescents exposed to significant early but not later childhood adversity. Dev Psychopathol. 2016;28(4pt2):1–15.
9. Meng H, Joyce AR, Adkins DE, Basu P, Jia Y, Li G, Sengupta TK, Zedler BK, Murrelle EL, van den Oord EJ. A statistical method for excluding non-variable CpG sites in high-throughput DNA methylation profiling. BMC Bioinformatics. 2010;11:227.
10. Farré P, Jones MJ, Meaney MJ, Emberly E, Turecki G, Kobor MS. Concordant and discordant DNA methylation signatures of aging in human blood and brain. Epigenetics Chromatin. 2015;8:19.
11. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
12. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):115.
13. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proc Natl Acad Sci U S A. 2010;107(21):9546–51.
14. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30(1):207–10.
15. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA, Feinberg AP. Increased methylation variation in epigenetic domains across cancer types. Nat Genet. 2011;43(8):768–75.
16. Lemire M, Zaidi SHE, Ban M, Ge B, Assi D, Germain M, Kassam I, Wang M, Zanke BW, Gagnon F, Morange PE, Trgout DA, Wells PS, Sawcer S, Gallinger S, Pastinen T, Hudson TJ. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. Nat Commun. 2015;6:6326.
17. Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, Daunay A, Busato F, Mein CA, Manfras B, Dias KRM, Bell CG, Tost J, Boehm BO, Beck S, Leslie RD. Identification of type 1 diabetes associated DNA methylation variable positions that precede disease diagnosis. PLoS Genet. 2011;7(9): e1002300.
18. Stringhini S, Polidoro S, Sacerdote C, Kelly RS, van Veldhoven K, Agnoli C, Grioni S, Tumino R, Giurdanella MC, Panico S, Mattiello A, Palli D, Masala G, Gallo V, Castagn R, Paccaud F, Campanella G, Chadeau-Hyam M, Vineis P. Life-course socioeconomic status and DNA methylation of genes regulating inflammation. Int J Epidemiol. 2015;44(4):1320–30.
19. Edgar R, Tan PPC, Portales-Casamar E, Pavlidis P. Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. Epigenetics Chromatin. 2014;7(1):28.
20. Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, Troakes C, Turecki G, O'Donovan MC, Schalkwyk LC, Bray NJ, Mill J. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. Nat Neurosci. 2016;19(1):48–54.
21. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, Monick M, Brody GH, Tan K, Beach SRH, Philibert RA. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.
22. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29(2):189–96.
23. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
24. Jones MJ, Islam SA, Edgar RD, Kobor MS. Adjusting for cell type composition in DNA methylation data using a regression-based approach. Methods Mol Biol. 2015. doi:10.1007/7651_2015_262.
25. Monick MM, Beach SRH, Plume J, Sears R, Gerrard M, Brody GH, Philibert RA. Coordinated changes in AHRR methylation in lymphoblasts and pulmonary macrophages from smokers. Am J Med Genet B Neuropsychiatr Genet. 2012;159B(2):141–51.
26. Philibert RA, Beach SRH, Lei MK, Brody GH. Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. Clin Epigenetics. 2013;5(1):19.
27. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, Belvisi MG, Brown R, Vineis P, Flanagan JM. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition. EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet. 2013;22(5):843–51.

28. Bauer M, Fink B, Thrmann L, Eszlinger M, Herberth G, Lehmann I. Tobacco smoking differently influences cell types of the innate and adaptive immune systemindications from CpG site methylation. Clin Epigenetics. 2016;8:83.
29. Tsai PC, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. Int J Epidemiol. 2015;44(4):1429–41.
30. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8(3):389–99.