OPEN

Citation: Transl Psychiatry (2017) 7, e1187; doi:10.1038/tp.2017.171

www.nature.com/tp

ORIGINAL ARTICLE

BECon: a tool for interpreting DNA methylation findings from blood in the context of brain

RD Edgar[1,2], MJ Jones[1,2], MJ Meaney[3,4,5], G Turecki[6] and MS Kobor[1,2,5]

Tissue differences are one of the largest contributors to variability in the human DNA methylome. Despite the tissue-specific nature of DNA methylation, the inaccessibility of human brain samples necessitates the frequent use of surrogate tissues such as blood, in studies of associations between DNA methylation and brain function and health. Results from studies of surrogate tissues in humans are difficult to interpret in this context, as the connection between blood–brain DNA methylation is tenuous and not well-documented. Here, we aimed to provide a resource to the community to aid interpretation of blood-based DNA methylation results in the context of brain tissue. We used paired samples from 16 individuals from three brain regions and whole blood, run on the Illumina 450 K Human Methylation Array to quantify the concordance of DNA methylation between tissues. From these data, we have made available metrics on: the variability of cytosine-phosphate-guanine dinucleotides (CpGs) in our blood and brain samples, the concordance of CpGs between blood and brain, and estimations of how strongly a CpG is affected by cell composition in both blood and brain through the web application BECon (Blood–Brain Epigenetic Concordance; https://redgar598.shinyapps.io/BECon/). We anticipate that BECon will enable biological interpretation of blood-based human DNA methylation results, in the context of brain.

Translational Psychiatry (2017) **7,** e1187; doi:10.1038/tp.2017.171; published online 1 August 2017

INTRODUCTION

Research exploring the associations and underlying mechanisms of complex traits such as brain function and health have primarily focused on genetic variation, with some success.[1–4] Inter-individual variation in brain function and health emerges as a result of both genetic variation and environmental influences. Enduring effects of environmental 'exposures' on brain function are of particular interest for our understanding of the origins of brain disorders and for the development of effective biomarkers. Epigenetic signals are an attractive candidate mediator of enduring environmental effects on cellular function. Indeed, there is now considerable evidence for the idea that environmentally regulated epigenetic states might form the biological basis for gene × environment interactions.[5–11] DNA methylation (DNAm) is a relatively stable epigenetic mark that is amenable to genome-wide assessment in biosamples from human subjects in studies of complex traits. The model of DNAm as a mediator of complex traits has produced a surge of DNAm-based, epigenome-wide association studies (EWAS) in brain research with promising results. Studies of mammalian models of stress, anxiety, addiction and brain cell composition support the hypothesis that DNA methylation patterns are associated with the brain function and health.[12–17] Moreover, there is evidence from human samples of specific patterns of DNAm linked to schizophrenia, autism, bipolar disorder and major psychosis.[18–22]

Tissue type is one of the strongest contributors to changes in methylation seen in EWAS,[23–25] and therefore an important consideration in designing EWAS. Blood is a commonly used surrogate for brain in human studies due to accessibility and potential for direct relation to disease through hormonal and immune regulation.[26–29] However, due to the highly tissue-specific nature of DNAm, it is important to consider the concordance between tissues to interpret findings derived from surrogate tissues.[23–25] Indeed, previous work has demonstrated that genome-wide DNAm profiles are highly tissue-specific, both in terms of inter-individual variability and absolute measures.[24,30–33] In blood and brain specifically, using a Illumina 450 K Human Methylation Array ((Illumina, San Diego, CA, USA, 450 K) data set of matched blood and brain tissues, we have shown that tissue identity, followed by cell-type heterogeneity within a tissue, represent the largest contributors to DNAm variance.[31] Moreover, while human tissues share some common DNAm patterns associated with biological variables like aging,[34] tissues also show unique patterns in relation to age.[31] Despite these findings, it remains largely unknown to what degree DNAm changes at individual cytosine-phosphate-guanine dinucleotides (CpGs) found in blood can serve as biologically relevant indicators of human brain biology.

To enable the interpretation of DNAm results from blood-based EWAS, we aimed to quantify to what degree blood is informative of brain DNAm. Using genome-wide analysis from paired human blood and brain run on the 450 K, we measured the level of concordance in DNAm across the methylome. Using variability and correlation thresholds on all CpGs, our results showed varying degrees of blood–brain concordance between CpGs. As we found concordance to be highly CpG dependent, blood-based studies do indeed need to be carefully interpreted in the context of tissues-specific DNAm differences. Therefore, as a tool for the community, we have built a user friendly web application Blood–Brain

[1]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada; [2]Centre for Molecular Medicine and Therapeutics, BC Children's Hospital, Vancouver, BC, Canada; [3]Ludmer Centre for Neuroinformatics and Mental Health, Douglas Mental Health University Institute, McGill University, Montreal, QC, Canada; [4]Singapore Institute for Clinical Sciences, Singapore; [5]Canadian Institute for Advanced Research, Toronto, ON, Canada and [6]Department of Psychiatry, McGill University, Montreal, QC, Canada. Correspondence: Dr MS Kobor, Centre for Molecular Medicine and Therapeutics, BC Children's Hospital, 950 West 28th Avenue, A5-162, 950 West 28th Avenue, Vancouver, BC V5Z 4H4, Canada. E-mail: msk@cmmt.ubc.ca

Received 24 February 2017; revised 14 June 2017; accepted 17 June 2017

Epigenetic Concordance (BECon) to explore blood-based DNAm findings in the context of human brain.

MATERIALS AND METHODS

Data collection

Using data from a previously published cohort, a total of 16 subjects were included in this study[31] (one subject from the original cohort of 17 did not have a blood sample and therefore could not be used; GSE95049).

Quality control and normalization of 450K data

Within Genome Studio samples were normalized by background subtraction and color correction, after which data were exported into R version 3.1.1. Sixty-five probes were removed as they directly measure single nucleotide polymorphisms (SNPs) and were not needed in this analysis beyond confirming replicate ID. Probes with evidence of cross-hybridization to regions other than the probe's target in the genome were removed (41 937 probes).[35] In addition, 1035 probes were filtered as no calls (bead count < 3) in 5% of samples, and 1342 probes were filtered as having 1% of samples with a detection *P*-value > 0.05. In total probe, filtering left 441 198 probes in the processed 450 K data set.

Normalization was performed using BMIQ,[36] as quantro[37] determined that quantile normalization would not be appropriate for this data set. The inappropriateness of quantile normalization was expected as our data set consisted of two very distinct tissues, with very different DNAm beta value distributions (Supplementary Figure S1).

The 63 samples (4 samples from 15 subjects and 3 samples from the one subject missing BA20), were examined with principal component analysis to visualize the presence of batch effects. In this study, the first principal component (PC) is not shown in visualizations (Supplementary Methods). The loadings of each PC were associated with technical and biological variables using ANOVA for categorical variables or Spearman correlations for continuous variables. Array barcode as well as refrigeration delay and sample pH showed strong associations with the top PCs loadings on samples (Supplementary Figure S2). ComBat was used to remove the batch effects of array barcode, refrigeration delay and PH from the DNAm data.[38]

Cell composition

As the data set was comprised of blood and brain samples, cell composition was estimated for each of the two tissue types. Specifically, blood cell-type proportions were estimated based on reference epigenomic profiles for six major sorted cell types[39] and normalized between blood samples.[40] Similarly, neuron and non-neuronal brain cell-type proportions were estimated[41] and normalized between brain samples[40] (Supplementary Figure S3).

As a check of the pre-processing steps, root mean squared errors were calculated between replicates. Root mean squared errors between replicate pairs remained high across all stages of quality control and pre-processing, indicating normalization and batch correction successfully removed noise from the data (Supplementary Figure S4).

Differential DNA methylation analysis

Mean DNAm across samples of a tissue were correlated between all tissue pairs using Spearman correlations as a DNA methylome-wide indicator of general sample similarity. To quantify the similarity of tissues at individual CpGs, differential DNAm analysis at each CpG was performed between all tissues pairs. Linear mixed effects models were run with covariates for subject gender and age, and to account for the paired structure of the samples, subject ID was included as a random effect in the model. Multiple test correction was done on the nominal *P*-values of each tissue-pair comparison, using Benjamini–Hochberg correction.[42]

Informative CpG selection

The correlation of DNAm level at each CpG was calculated between blood and brain separately for each brain region using Spearman correlations on *M* values. The variability of a CpG across individuals was measured as the range between the 10th percentile and the 90th percentile of blood sample CpG betas.[43] This reference range is intended to capture variability in the bulk of the samples while limiting the effect of outlier measures at a CpG, which would otherwise give a falsely high estimate of variability.

To define 'informative CpGs', we used biologically relevant thresholds for both correlation and variability. The most highly correlated and variable CpGs observed were the polymorphic CpGs (CpGs with a known SNP at the cytosine or guanine of the CpG) and those on the X and Y chromosomes (presumably highly variable because the cohort contains males and females). Although these 32 344 CpGs were not of explicit interest in this study, and later removed from analysis, the variability and correlation distributions of these 32 344 highly correlated polymorphic and sex chromosome CpGs were used to guide the selection of the correlation and variability thresholds for informative CpGs. A full explanation of the threshold selection is provided in the Supplementary Text, but in short, informative CpGs had to meet a variability threshold of at least a 0.1 difference between 90th and 10th percentile of blood CpG beta values and correlation threshold of 2 standard deviations from the mean correlation of the highly positively correlated polymorphic and sex chromosome CpGs-enriched correlation peak, defined separately for each brain region.

The importance of a variability threshold was clear when the matched blood and brain samples were randomly unmatched in five simulations. The correlation distributions of these unmatched data sets were used as null distributions to compare with the real paired data correlation distributions.[44]

CpG to gene associations

There are multiple approaches for associating a CpG to a gene, such as the closest transcription start site,[35] associating CpGs to gene by the CpGs localization to a gene's body or promoter,[45] or stringent associations based on CpGs with only one likely gene association (that is, lone gene associations).[31] These approaches focus on a single gene, rather than allowing for multiple gene associations for a CpG. We used a CpG to gene association definition that allows for a CpG to be associated with multiple gene features, as well as multiple genes (Supplementary Text). This inclusive association, while somewhat more ambiguous, is an attempt to capture all possible roles of a CpG in gene regulation.[46] The gene list used was the Refseq genes from UCSC, including all splice variants of Refseq mRNA (Supplementary Text). The gene list included 24 047 genes and a total of 33 431 unique transcription units. The 485 512 CpGs on the 450 K array associated with 23 018 genes (43.8% intragenic CpGs, 34.2% promoter CpGs, 2.5% 3′ region CpGs, and 19.5% intergenic CpGs).[46]

Comparison to previous analyses

DNAm in blood and brain samples was analyzed previously using a 450K data set[33] providing an opportunity to validate our findings using an independent data set. The published data set provided 74-matched brain and blood samples on Gene Expression Omnibus (GEO; GSE59685).[33] The regions examined in this previous work were cerebellum, entorhinal cortex, frontal cortex and superior temporal gyrus regions. Although their frontal cortex and our BA10 region partially overlap, the three other regions available in GSE59685 allow for possible validation in structurally and functionally different brain regions. We ran the GSE59685 normalized data through our pipeline (described above) to make the results as comparable as possible. Unfortunately, in GSE59685, the tissues were run on separate arrays, introducing confounding of array and tissue. However, despite this limitation, to be consistent with our analysis, we did run ComBat to correct for sentrix ID. Cell correction was performed as described above in the brain regions and blood. Spearman correlations and reference ranges were calculated between blood and all brain regions, and informative CpGs were defined similarly as described above. The actual percent overlap of informative CpGs was calculated for all seven brain regions available. Monte Carlo simulations were used to build an expectation of overlap between two lists of informative CpGs.

To test for enrichment of DNAm quantitative trait loci (mQTL) associations in informative CpG lists, we used mQTL previously identified at $P < 1 \times 10^{-14}$ in middle aged individuals using the mQTL database[47] (http://www.mqtldb.org/). The mQTL list contained 31 325 CpGs under observed genetic influence. Using Monte Carlo simulations, we built an expected overlap of the mQTL-associated CpGs and informative CpGs, to compare with the observed overlap.

There have been numerous studies using blood as a surrogate for brain when studying a neurobiological disorder.[48] To explore whether CpGs identified in these studies are informative of brain DNAm, we collected a list of six CpGs associated with four genes previously observed to be differentially methylated in blood in relation to psychiatric disorders.[48] We explored the identified CpG's correlation between blood and brain in

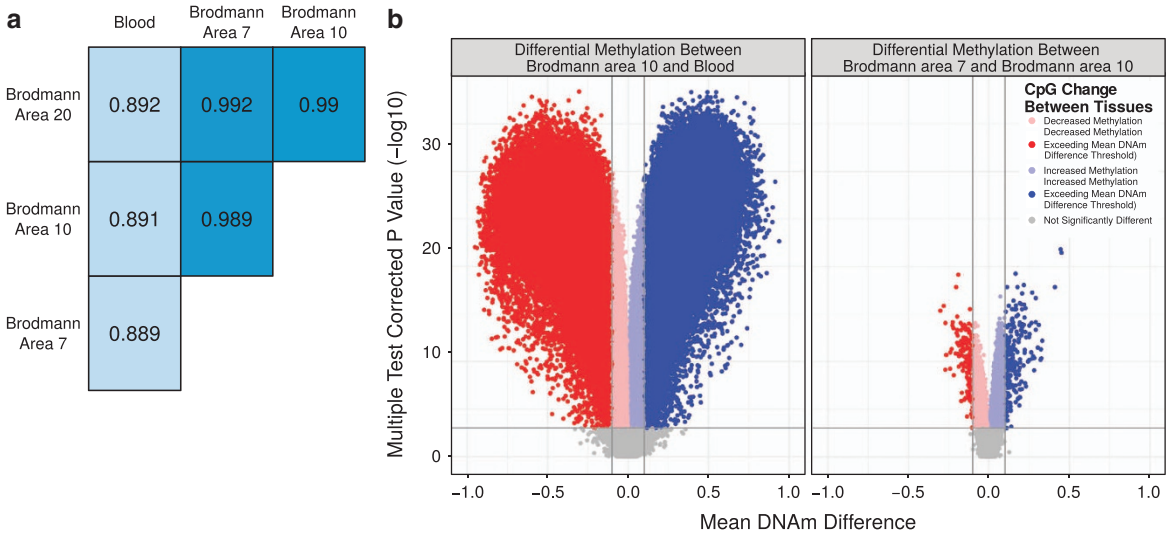

Figure 1. Human blood and brain show very distinct methylation patterns. (**a**) DNA methylation (DNAm) correlation values between each tissue-pair from an individual, averaged across all individuals. (**b**) Volcano plots of the differential methylation analysis between representative tissue pairs (blood and Brodmann area 10; and Brodmann area 10 and 7). Vertical lines indicate a DNAm difference between compared tissues of 0.1. The horizontal line represents an false discovery rate (FDR)-corrected *P*-value of 0.001. Points are colored to highlight cytosine-phosphate-guanine dinucleotides (CpGs) exceeding both the biological and statistical cutoffs.

BECon. Specifically, the CpGs were evaluated in terms of the correlation percentile in each brain region to explore if the CpGs are more or less informative then the average CpG. CpGs we also examined for variability in each tissue as a measure of biological relevance.

Code availability

We have made the code used for the entire analysis, along with BECon available on GitHub (github.com/redgar598/BECon).

RESULTS

Differential methylation analysis of blood and brain

Tissue is the one of the largest contributor to DNAm variance.[31,49] However, it is not yet known which specific CpGs show concordance in DNAm between blood and brain, at which blood DNAm could potentially serve as a proxy for brain DNAm, and which CpGs show no concordance. DNAm was measured from matched human blood and brain on the 450K.[45] In comparisons of DNAm between the four samples from each of the 16 individuals across the methylome, different brain regions from the same individual had higher correlation with each other than any brain region with blood (Figure 1a). In addition, brain to blood DNAm analysis returned orders of magnitudes more differentially methylated CpGs than did differential DNAm analysis between brain regions (for example, 120 970 differential CpGs between BA10:blood and 459 between BA10:BA7, false discovery rate < 0.001, mean difference in DNA methylation between tissues 0.1; Figure 1b).

Informative CpGs exist between human blood and brain

Although our group and others have observed large differences between human blood and brain DNAm,[23,31] by necessity, blood is often used as a surrogate for brain tissue. We therefore set out to use the strength of our matched sample cohort to identify CpGs that show concordance between blood and brain. Our first step in identifying these CpGs was to use the correlation of inter-individual variability between blood and each brain region (Figure 2a; Supplementary Figures S5–S7). These correlations had a slightly skewed distribution toward negative correlations, indicating the majority of CpGs are not concordant between blood and brain, but a few CpGs are highly positively correlated (Supplementary Figure S8; Supplementary Text).

However, many of the most highly correlated CpGs had very low variability between individuals, which likely limits the utility of these CpGs to explain differences in phenotype and/or exposures in EWAS. We therefore explored the importance of variability in defining concordant CpGs. We used reference range as a measure of variability to limit the impact of outlier samples and non-normally distributed methylation values at individual CpGs, which is especially important in our relatively small cohort. We looked at the correlation distributions of increasingly more variable sets of CpGs (Figure 2b) and found that the higher the variability threshold, the more skewed to positive correlations the distribution became. This trend of skewing toward positive correlations was not an artifact of the variability measurements as it disappeared when the data were simulated as unpaired (Figure 2b). We therefore endeavored to select CpGs with both high inter-individual variability, as well as high blood–brain correlation.

To make the thresholds of variability and correlation more biologically driven, we based the thresholds on a set of CpGs which were some of the most highly variable and correlated between blood and brain, polymorphic CpGs and CpGs on sex chromosomes (Figure 3a). We focused on variability in blood and blood–brain correlation of these 32 344 CpGs to define our thresholds. The reference range variability of these CpGs in blood was 0.11, so a threshold of 0.1 was used for our concordant CpG selection (Figure 3b). The correlation distribution of all CpGs was bimodial, and we focused on the polymorphic and sex chromosome enriched highly positively correlated peak to define our correlation thresholds (Figure 3c). Therefore, our definition of a blood–brain 'informative CpG' is a CpG at which the DNAm in blood correlated with DNAm in brain (absolute r_s = BA7 0.36; BA10 0.40; BA20 0.33) and DNAm is also highly variable in blood (reference range > 0.1).

Using our variability threshold of 0.1, we identified 83 427 variable CpGs. Of these, 48% also passed our correlation requirements, resulting in a total of 40 029 informative CpGs (both positively and negatively correlated). Thus, 9.7% of the total number of CpGs examined were informative between blood and any of the three brain regions. Informative CpGs identified in each

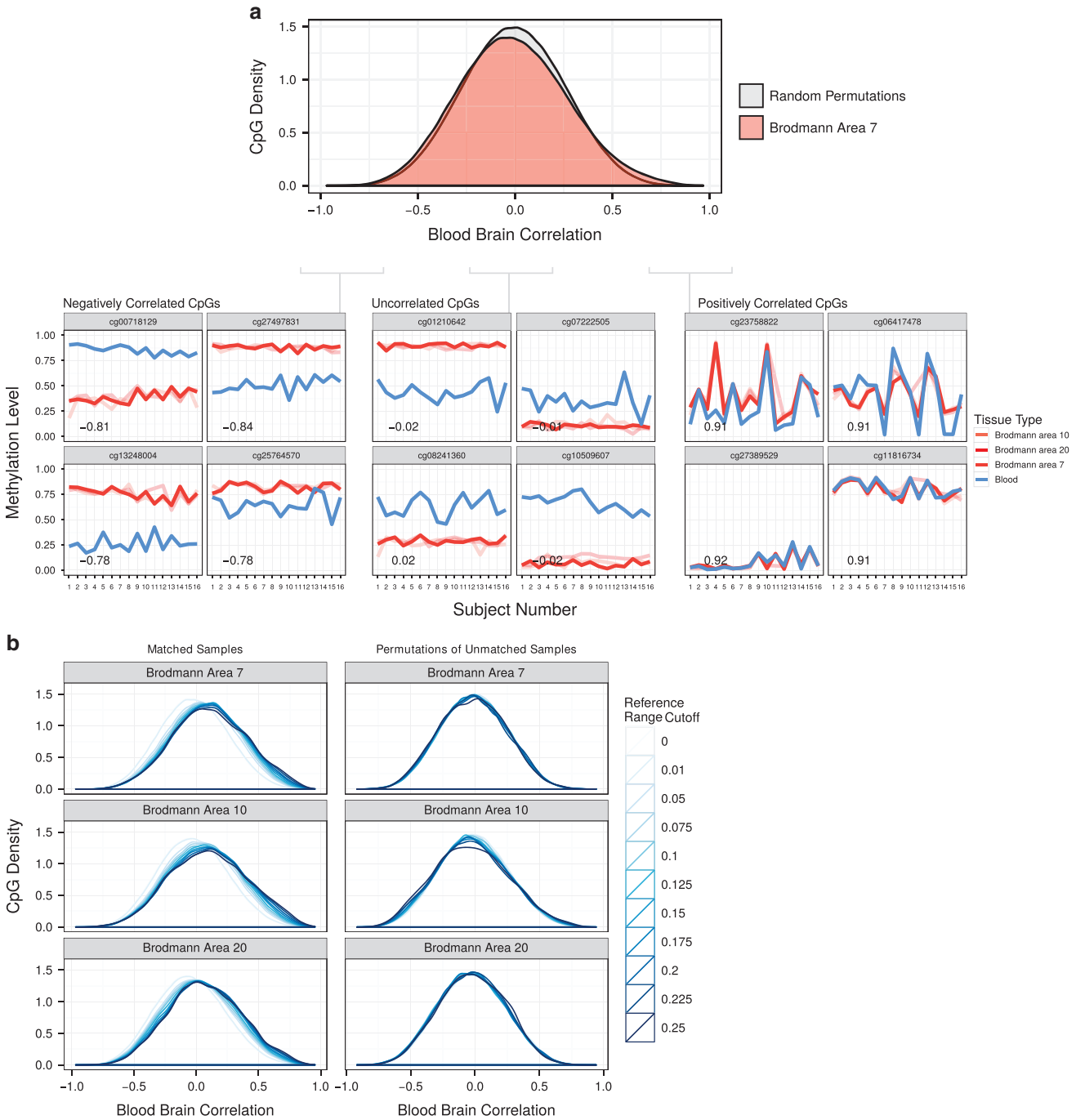

Figure 2. Paired tissues allow the identification of blood–brain concordance across individuals. (**a**) Representative distribution of all CpG's correlation values between blood and Brodmann area 7. The filled gray density curves show the random permutations of unpaired samples, whereas filled red curves show the correlation distribution of the paired samples. Line plots show DNA methylation (DNAm) of representative CpGs from three sections of the correlation distribution. The line plots show the correlation of inter-individual variability at the representative CpGs. Lines are colored by tissue and BA10 and BA20 are shown fainter. Spearman correlations are shown for the correlation between blood DNAm and Brodmann area 7 DNAm. (**b**) Correlation distributions of CpGs are shown for each brain region at increasingly stringent variability cutoffs. Line colors darken as the CpGs underlying the distribution become more strictly thresholded on reference range. Plots on the left show the distributions of matched blood and brain correlation values, and on the right the distributions are for unmatched permutations of sample order to simulate unpaired data. CpG, cytosine-phosphate-guanine dinucleotide.

brain region show a large overlap (Figures 3d and e), as expected considering the observed similarities in DNAm of the three brain regions. Although we have been strict in our definition of what an informative CpG is, and small changes to the correlation and variability thresholds do result in large changes for the number of CpGs considered informative (Supplementary Table S1). Our informative CpGs were the most informative by our criteria and relative to the rest of the 450K CpGs. We note, however that blood–brain DNAm was rarely very highly correlated when applying more stringent criteria (313 total CpGs, from any brain region, exceeded an absolute $r_s = 0.80$ and reference range 0.1; Supplementary Table S1). Although at the correlation values, we

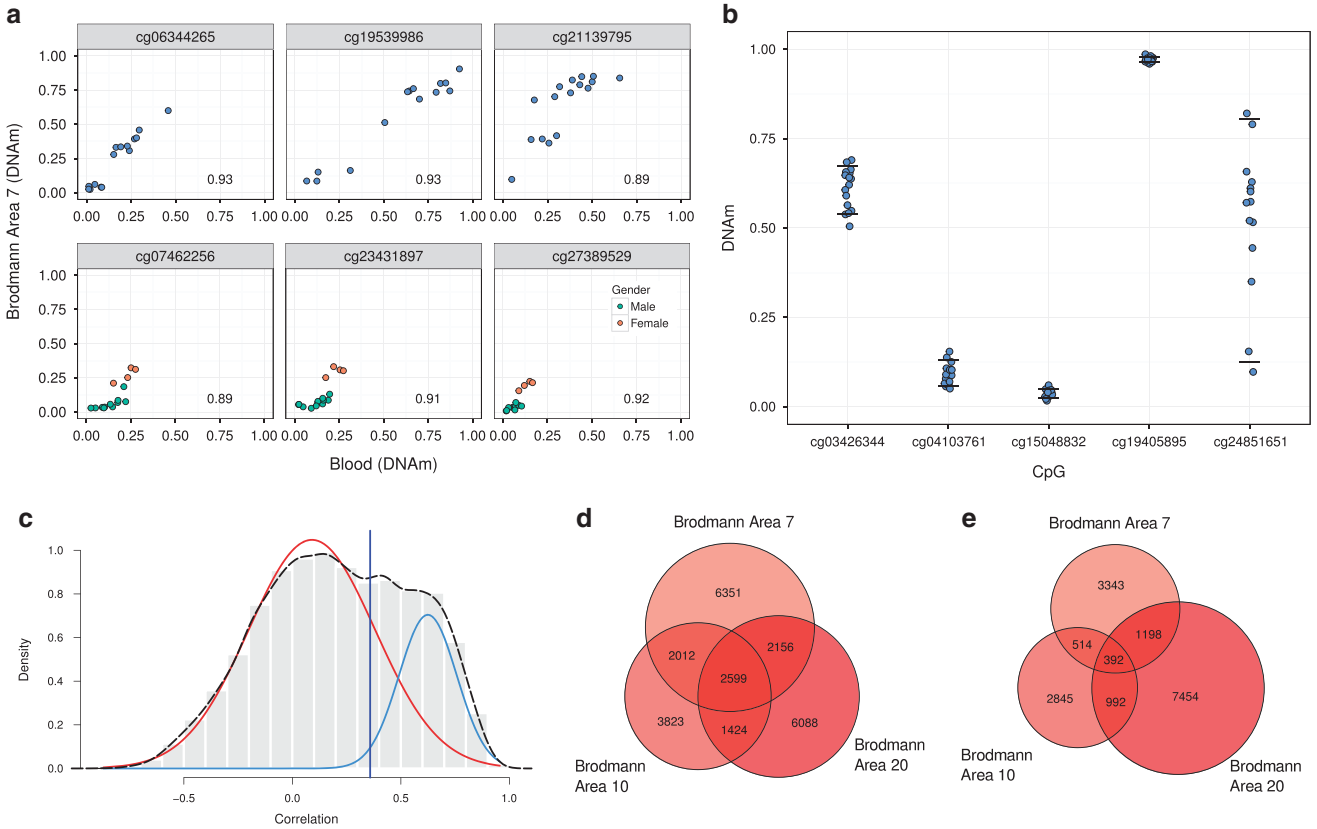

Figure 3. Informative cytosine-phosphate-guanine dinucleotides (CpG) are defined through both variability and correlation thresholds. (**a**) Representative single nucleotide polymorphism (SNP) and sex chromosome CpGs show high levels of correlation and variability. All plots show the relationship between DNAm in Brodmann area 7 and blood with the correlation coefficient in the bottom right of the plot. The CpGs in the top three plots show a polymorphic CpG. The CpGs in the bottom three plots show sex chromosome CpGs, and individuals are colored by gender. (**b**) Reference range variability in blood DNAm is shown at CpGs representative of the spectrum of variability seen at 450K CpGs. Horizontal lines at each CpG represent the 90th and 10th percentile of blood DNAm level. Reference range is the range between the horizontal lines. (**c**) Definition of the correlation coefficient threshold for informative CpGs in Brodmann brain area 7. The histogram and broken line show the correlation distribution for CpGs passing the strictest variability threshold. Solid lines are the two fitted Gaussian components of the distribution (red generally uncorrelated peak, blue positively correlated peak). The vertical black line indicates two standard deviation away from the positively correlated peak mean, which was used as the correlation threshold for blood–brain informative CpGs. (**d**) Venn diagram showing the overlap of informative CpGs with positive correlations between blood and brain in the three Brodmann areas sampled. (**e**) Venn diagram showing the overlap of informative CpGs with negative correlations between blood and brain in the three Brodmann areas sampled.

have used our informative CpGs would have variability in DNAm between blood and brain, our informative list has been built on biologically defined statistical thresholds based on the polymorphic and sex chromosome CpGs characteristics. Therefore, our informative CpG list is relatively stringent and should reflect the strongest concordance signal in the data.

Informative CpGs were enriched in intergenic regions of the genome

Next, we explored the genomic location(s) of our informative CpGs. We associated each CpG with a gene (Supplementary Text), and used the Illumina annotation for CpG island associations.[45] In general, informative CpGs were depleted in gene promoters and CpG islands and enriched in intergenic regions ($P < 0.001$; Figure 4).

We then explored whether genes associated with informative CpGs were involved in any specific biological process. We used a list of 239 genes associated with at least ten informative CpGs (informative genes), to focus on the genes with high DNAm variability and concordance between blood and brain. The informative gene list showed enrichment for Gene Ontology terms related to cell-adhesion and highly multifunctional genes (Supplementary Table S2).[50] We then investigated whether the informative genes were more highly expressed in either blood or brain, using independent gene expression data sets (GSE17612, GSE37171 and GSE61635). The informative genes are less expressed in blood samples than the average expression of all genes measured ($P < 0.001$, Wilcoxon-rank sum) but the expression of the informative genes was not different from the average expression of all genes in brain samples ($P = 0.98$, Wilcoxon-rank sum; see Supplementary Text and Supplementary Figure S9). Therefore, informative genes may be more brain-specific than blood specific as they are less expressed in blood then the average gene.

Comparison of informative CpGs to previous findings

An existing similar blood–brain DNAm analysis[33] provided an opportunity for independent validation of our results. The previous study, reports the correlation between blood and four brain regions (cerebellum, entorhinal cortex, frontal cortex and superior temporal gyrus), and provides data for 74 individuals with paired blood samples DNAm (GSE59685). We found that although the greatest overlap of informative CpGs was between brain regions from the same study, the overlap between the lists of

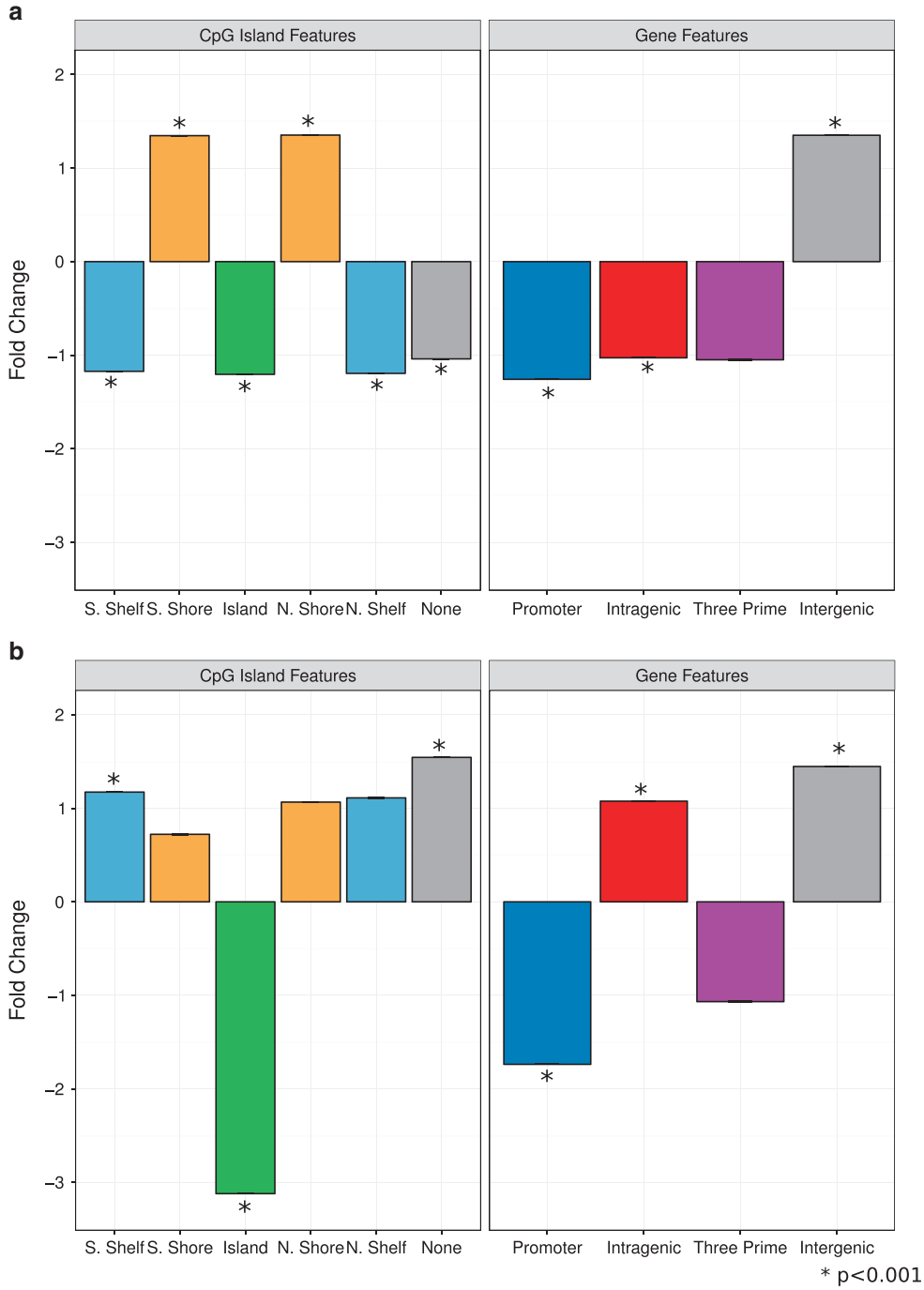

Figure 4. Blood brain informative cytosine-phosphate-guanine dinucleotides (CpGs) show associations to specific genomic features. In all plots, bars show the fold-change between informative CpG count in each region and the mean count of randomly selected CpGs in that same region, from 10 000 iterations. Error bars show standard error. (**a**) Genomic enrichment for informative CpGs, which are positively correlated between blood and brain. (**b**) Genomic enrichment for informative CpGs, which are negatively correlated between blood and any brain.

informative CpGs from the two studies was greater than expected by chance (Figure 5a). Interestingly, the previous study had very few negatively correlated sites (42 negatively correlated informative CpGs compared to 16 738 in our data set), which may suggest the negative correlations we observed represent a property inherent to our smaller sample size (Supplementary Figure S10).

As polymorphic CpGs were some of the most highly concordant CpGs in our data, we speculated that our informative CpGs could be under genetic influence and represent potential mQTL. We tested whether our informative CpGs and those we defined in GSE59685 were enriched for associations to known mQTL using the mQTL database[47] (http://www.mqtldb.org/). We found 8202 out of 40 029 (21%) of our informative CpGs and 3018 out of 10 930 (28%) informative CpGs from the Hannon *et al.*[33] cohort were previously identified as associated with an mQTL. In both informative lists, the mQTL-associated CpG numbers represented significant enrichment for known mQTL associations (Monte Carlo simulations, $P < 0.0001$). Although our smaller cohort, likely, did not have enough genetic diversity to capture all potential mQTL

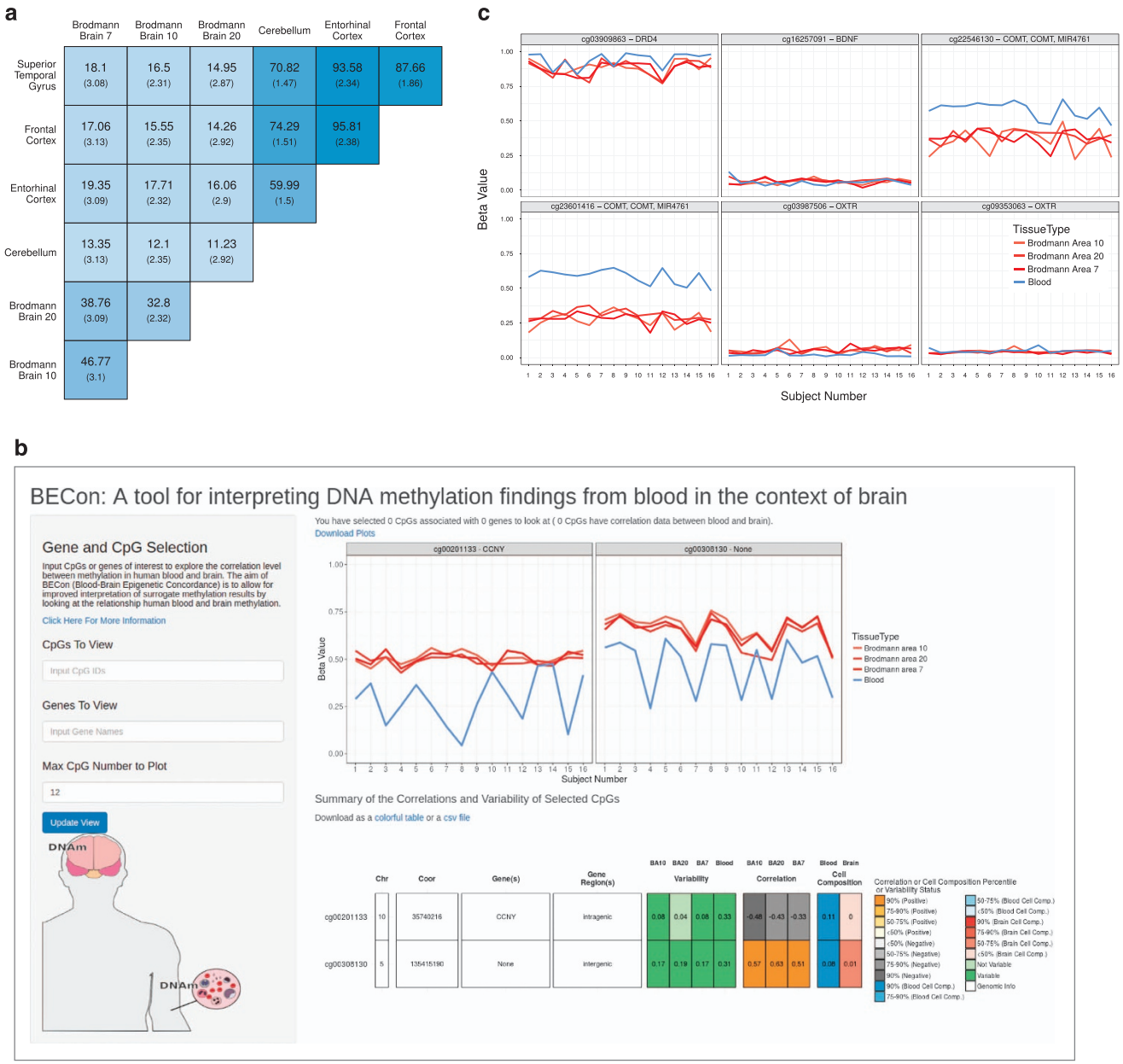

Figure 5. The informative cytosine-phosphate-guanine dinucleotide (CpG) list can be used to validate previous findings. (**a**) Percent of overlapping positively correlated informative CpGs in each of our brain regions and those regions used in the Hannon *et al.*[33] data. The number in larger text is the percent of the informative CpGs along the rows that are also informative in the tissue along the columns, the smaller number is the percent overlap of these lists expected by chance. The color of each box is based on the percent overlap. (**b**) Visualization provided in BECon. The plots show the inter-individual variability at two representative CpGs. The table shows the various metrics provided in BECon for each CpGs queried. (**c**) Inter-individual variability at the CpGs identified previously in blood-based studies of psychiatric disorders.

sites (Supplementary Figure S11), our informative CpGs were still enriched for mQTL associations.

BECon as a community resource

The availability of this matched blood and brain DNAm data set provided an excellent opportunity to develop a community resource. We have built an R Shiny web application[51] to aid interpretation of blood-based DNAm results in studies of brain function and health (Blood–brain Epigenetic Condordance; BECon; https://redgar598.shinyapps.io/BECon/). Although we also made the full data set available on GEO (GSE95049), we built BECon for researchers interested in a particular gene or CpG, but without the need to re-analyze our data themselves. A full description of the information provided through BECon is in the Supplementary Text, with detailed explanations for how the metrics provided were calculated. To aid interpretation of epigenetic results, BECon provides metrics on the variability of CpGs in our 16 matched blood and brain samples, concordance DNAm at CpGs between blood and brain, and estimations of how strongly a CpG is affected by cell composition in both blood and brain (Figure 5b).

To demonstrate the utility of BECon, we assessed key candidate genes often investigated for changes in DNAm in relation to psychiatric disorders (BDNF, COMT, OXTR and DRD4). We examined six specific CpGs in these candidate genes that were identified as differentially methylated in blood in studies of

psychiatric disorders[48] (Supplementary Table S3). The six CpGs show varying levels of average correlation across brain regions ($r_s = -0.15$ to 0.49) and one CpG is in the 90th percentile of all CpG correlation values (Figure 5c; Supplementary Table S4), suggesting only some differential DNAm reported previously in blood could be expected to be seen in brain.

DISCUSSION

The increasing popularity of EWAS using blood samples to study brain function and health outcomes has created a need for tools to enable interpretation of DNAm results in the context of the brain. Our findings indicate that it is essential to examine the concordance of DNAm between blood and brain at each CpG before interpreting blood-based results, as concordance varied greatly dependent on CpG. A subset of CpGs, which we consider informative of brain, has been validated in an independent cohort. Despite the tissue-specific DNAm seen between blood and brain, the validated informative CpGs suggested blood has applicability as a surrogate for brain. In identifying informative CpGs, correlation seemed to be the most logical measure of concordance; however, we found a variability measure was also necessary to identify concordance with the utility to explain differences in phenotype and/or exposures in EWAS. Discordant CpGs were either not variable in one tissue, or appear to be potentially tissue-specific in their variability. Tools to examine multiple metrics of concordance simultaneously will aid the interpretation of blood-based DNAm results. We developed BECon to enable easier examination of the concordance of blood and brain DNAm. Our informative CpGs were the most concordant relative compared with the rest of the 450K. It is important to consider with blood–brain correlations of 0.33–0.40, there will be a substantial variance in the brain DNAm when using blood as a surrogate, even at our informative CpGs. Not surprisingly, using a stricter correlation threshold ($r_s = 0.80$) few CpGs would be considered concordant (Supplementary Table S1). Examining concordance of individual CpGs of interest, through BECon, might be a more useful and meaningful representation of the data for some researchers than focusing on the subset of the 450K, we have categorized as informative. Regardless, our hope is that BECon will enable biologically grounded interpretation of blood-based DNAm results genome-wide. Our hope is that BECon that should allow for biologically grounded interpretation of blood-based DNAm results genome-wide.

The web application BECon that we provide to the community includes metrics on the variability of CpGs in blood and brain. We have included metrics on variability as we found that the DNAm at many correlated CpGs varied by less that would generally be considered biologically meaningful. It is the consensus of the field that differences in DNAm, whether between tissues or disease states, to be have a theoretical biological impact they must show variability in DNAm measures.[52] When examining concordance of DNAm in BECon, it is expected that if a CpG shows low variability in either blood or brain tissue, that any concordance will not be biologically meaningful. Although CpGs not variable in our blood and brain data set may vary in another context, or perhaps another tissue, we are hopeful that or concordance findings are robust and relevant to other blood and brain samples.

We have some evidence our concordance findings are robust to brain-region type and cohort as we were able to confirm the general trends seen in previous blood–brain studies.[33,53] Our study, like others, demonstrated that tissue-type is a considerable contributor to DNAm variability, evident by the abundant DNAm differences we have observed between each brain region and blood. Second to broad tissue-type differences, the next largest contributor to DNAm variation is cell composition within a tissue.[31] In studies using surrogate tissues, as with all DNAm studies, it has become more apparent that cell-type composition adjustments are a mandatory step in the analysis of data for results to be interpretable. We therefore included statistics on the estimated effect of cell-type composition in BECon to better enable researchers to examine the effect of cell composition on CpGs or gene of interest.

Next to tissue-type and cell composition, genetics are a major contributor to DNAm variability. CpGs under the influence of SNPs are some of the most variable CpGs observed in DNAm.[47,54,55] Given our variability selection criteria, it is not surprising that out informative CpGs were enriched for associations to mQTL. However, it is reasonable to suspect that, regardless of variability criteria, CpGs under genetic control will be the most concordant between blood and brain as the genetics will be consistent between tissues. When interpreting DNAm concordance between blood and brain through BECon it will be important to consider the biology driving the concordance. At some CpGs, the driver of variability and concordance may be primarily genetic and potentially independent of any environmental influences seen in associated with blood DNAm. Although we were able to observe enrichment for associations to mQTL, we speculate that due to our smaller sample size we were not as sensitive to detect mQTL associations as the previous blood brain analysis.[33] With 16 individuals, we may not have had representative samples from all possible alleles at many potential mQTL SNP loci. Therefore, DNAm at potential mQTL-associated CpGs was less variable in our data and high correlations could not be observed. Interestingly, this suggests that there is a minimum required sample size for thorough mQTL detection above 16 individuals; however, we can not speculate if the 74 individuals used in previously were enough to detect all possible mQTL. Likely, the detection power of mQTL will scale with sample size, as previously seen with gene expression QTL detection by the GTEx Consortium.[56]

Future studies into the gene regulation and expression associations of concordant CpGs in blood and brain may provide insight into the biological relevance of blood DNAm in brain. Although we were unable to directly compare the concordance of our sites in gene expression data, we were able to look at the overall expression of our informative genes in brain and blood. Interestingly, we found that informative genes are less expressed in blood than expected by chance. It is possible that by our selection criteria, we have identified CpGs that can serve as biomarkers for brain-specific genes that have little to no function in blood and are therefore not highly expressed.

In addition to the highly tissue-specific nature of DNAm, using a surrogate tissue for brain DNAm is further complicated by the existence of higher levels of hydroxymethylation in the brain compared with other tissues.[57–59] Hydroxymethylation has been seen in the brain at levels as high as 0.65% but only at 0.027% in blood.[58,60] Here, we have characterized the anticipated utility of blood as a surrogate for brain in terms of a composite DNAm signal (mC+hmC). Hydroxymethylation in the human brain has potentially added complexity to our data and to the correlations calculated between blood and brain.

Using BECon, we were able to identify a CpG showing high concordance between blood and brain that has also been identified as differentially methylated in a study of blood from patients with schizophrenia and controls (cg03909863).[29] The CpG is a promising candidate to show DNAm concordance in the brains of the individuals, and could be relevant functionally in the brain as the CpG in located in coding region of dopamine receptor D4 (*DRD4*). In future studies that use blood as a surrogate for brain, BECon will enable prioritization of CpGs for validation to those CpGs with demonstrated concordance between blood and brain.

Despite the limitations of our analysis, there is a tremendous value and information content in quantifying the genome-wide concordance of DNAm in the blood and brain. In anticipation of the community's interest in examining whether specific CpGs in

blood are informative of brain DNAm, we have built BECon to enable examination of concordance between blood and brain. In addition, we have made the code used for analysis publicly available, which might be useful to the community as more blood–brain data become available, including these ascertained on higher resolution platforms like the Illumina Infinium Methylation EPIC array (Illumina). We expect BECon to be most useful to users who need to interpret blood DNAm results from a study of brain function and health. This application may also help guide blood-based surrogate studies toward candidate gene approaches or *post hoc* selection of CpGs for validation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

We thank Dr Magda Price, Sarah Goodman, Sumaiya Islam and Jack Hickmott for comments on the analysis and manuscript. This work was supported by: R Howard Webster Foundation (F13-00031 to MSK); W Garfield Weston Foundation/Brain Canada Foundation (F13-02369 to MSK, MJM and GT). MSK is a Canada Research Chair in Social Epigenetics and the BC Leadership Chair in Child Development.

REFERENCES

1 Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**: 441–447.
2 Tansey KE, Rees E, Linden DE, Ripke S, Chambert KD, Moran JL *et al.* Common alleles contribute to schizophrenia in CNV carriers. *Mol Psychiatry* 2015; **21**: 1085–1089.
3 Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* 2016; **530**: 177–183.
4 Day FR, Helgason H, Chasman DI, Rose LM, Loh PR, Scott RA *et al.* Physical and neurobehavioral determinants of reproductive onset and success. *Nat Genet* 2016; **48**: 617–623.
5 Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* 2009; **462**: 868–874.
6 Meaney MJ. Epigenetics and the biological definition of gene x environment interactions. *Child Dev* 2010; **81**: 41–79.
7 Klengel T, Mehta D, Anacker C, Rex-Haffner M, Pruessner JC, Pariante CM *et al.* Allele-specific FKBP5 DNA demethylation mediates gene-childhood trauma interactions. *Nat Neurosci* 2012; **16**: 33–41.
8 Beach SRH, Brody GH, Lei MK, Kim S, Cui J, Philibert RA *et al.* Is serotonin transporter genotype associated with epigenetic susceptibility or vulnerability? Examination of the impact of socioeconomic status risk on African American youth. *Dev Psychopathol* 2014; **26**: 289–304.
9 Klengel T, Binder EB. FKBP5 allele-specific epigenetic modification in gene by environment interaction. *Neuropsychopharmacology* 2015; **40**: 244–246.
10 Boyce WT, Kobor MS. Development and the epigenome: the 'synapse' of gene-environment interplay. *Dev Sci* 2015; **18**: 1–23.
11 Chen L, Pan H, Tuan TA, Teh AL, MacIsaac JL, Mah SM *et al.* Brain-derived neurotrophic factor (BDNF) Val66Met polymorphism influences the association of the methylome with maternal anxiety and neonatal brain volumes. *Dev Psychopathol* 2015; **27**: 137–150.
12 Blaze J, Scheuing L, Roth TL. Differential methylation of genes in the medial prefrontal cortex of developing and adult rats following exposure to maltreatment or nurturing care during infancy. *Dev Neurosci* 2013; **35**: 306–316.
13 Mancino S, Burokas A, Gutiérrez-Cuesta J, Gutiérrez-Martos M, Martín-García E, Pucci M *et al.* Epigenetic and proteomic expression changes promoted by eating addictive-like behavior. *Neuropsychopharmacology* 2015; **40**: 2788–2800.
14 Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S *et al.* Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron* 2015; **86**: 1369–1384.
15 Elliott E, Manashirov S, Zwang R, Gil S, Tsoory M, Shemesh Y *et al.* Dnmt3a in the medial prefrontal cortex regulates anxiety-like behavior in adult mice. *J Neurosci* 2016; **36**: 730–740.
16 Mychasiuk R, Muhammad A, Kolb B. Chronic stress induces persistent changes in global DNA methylation and gene expression in the medial prefrontal cortex, orbitofrontal cortex, and hippocampus. *Neuroscience* 2016; **322**: 489–499.
17 Saunderson EA, Spiers H, Mifsud KR, Gutierrez-Mecinas M, Trollope AF, Shaikh A *et al.* Stress-induced gene expression and behavior are controlled by DNA methylation and methyl donor availability in the dentate gyrus. *Proc Natl Acad Sci* 2016; **113**: 4830–4835.
18 Mill J, Tang T, Kaminsky Z, Khare T, Yazdanpanah S, Bouchard L *et al.* Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. *Am J Hum Genet* 2008; **82**: 696–711.
19 Kaminsky Z, Tochigi M, Jia P, Pal M, Mill J, Kwan A *et al.* A multi-tissue analysis identifies HLA complex group 9 gene methylation differences in bipolar disorder. *Mol Psychiatry* 2012; **17**: 728–740.
20 Pidsley R, Viana J, Hannon E, Spiers H, Troakes C, Al-Saraj S *et al.* Methylomic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia. *Genome Biol* 2014; **15**: 483.
21 Ladd-Acosta C, Hansen KD, Briem E, Fallin MD, Kaufmann WE, Feinberg AP. Common DNA methylation alterations in multiple brain regions in autism. *Mol Psychiatry* 2014; **19**: 862–871.
22 Jaffe AE, Gao Y, Deep-Soboslay A, Tao R, Hyde TM, Weinberger DR *et al.* Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* 2016; **19**: 40–47.
23 Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MPM, van Eijk K *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol* 2012; **13**: R97.
24 Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* 2013; **500**: 477–481.
25 Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518**: 317–330.
26 Fuchikami M, Morinobu S, Segawa M, Okamoto Y, Yamawaki S, Ozaki N *et al.* DNA methylation profiles of the brain-derived neurotrophic factor (BDNF) gene as a potent diagnostic biomarker in major depression. *PLoS ONE* 2011; **6**: e23881.
27 Unternaehrer E, Luers P, Mill J, Dempster E, Meyer AH, Staehli S *et al.* Dynamic changes in DNA methylation of stress-associated genes (OXTR, BDNF) after acute psychosocial stress. *Transl Psychiatry* 2012; **2**: e150.
28 Melas PA, Rogdaki M, ÖU Schalling M, Lavebratt C, Ekström TJ. Epigenetic aberrations in leukocytes of patients with schizophrenia: association of global DNA methylation with antipsychotic drug treatment and disease onset. *FASEB J* 2012; **26**: 2712–2718.
29 Cheng J, Wang Y, Zhou K, Wang L, Li J, Zhuang Q *et al.* Male-specific association between dopamine receptor D4 gene methylation and schizophrenia. *PLoS ONE* 2014; **9**: e89128.
30 Davies MN, Volta M, Pidsley R, Lunnon K, Dixit A, Lovestone S *et al.* Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol* 2012; **13**: R43.
31 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. *Epigenet Chromatin* 2015; **8**: 19.
32 Jiang R, Jones MJ, Chen E, Neumann SM, Fraser HB, Miller GE *et al.* Discordance of DNA methylation variance between two accessible human tissues. *Sci Rep* 2015; **5**: 8257.
33 Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics* 2015; **10**: 1024–1032.
34 Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013; **14**: R115.
35 Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenet Chromatin* 2013; **6**: 4.
36 Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2013; **29**: 189–196.
37 Hicks SC, Irizarry RA. When to use quantile normalization? *bioRxiv* 2014. Available at http://dx.doi.org/10.1101/012203.
38 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; **8**: 118–127.
39 Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* 2014; **15**: R31.
40 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* 2017; **1589**: 99–106.
41 Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* 2013; **8**: 290–302.
42 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995; **57**: 289–300.

43 Lemire M, Zaidi SHE, Ban M, Ge B, Aïssi D, Germain M *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat Commun* 2015; **6**: 6326.
44 Davies MN, Lawn S, Whatley S, Fernandes C, Williams RW, Schalkwyk LC. To what extent is blood a reasonable surrogate for brain in gene expression studies: estimation from mouse hippocampus and spleen. *Front Neurosci* 2009; **3**: 54.
45 Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* 2011; **98**: 288–295.
46 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. *Epigenet Chromatin* 2014; **7**: 28.
47 Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biol* 2016; **17**: 61.
48 Klengel T, Pape J, Binder EB, Mehta D. The role of DNA methylation in stress-related psychiatric disorders. *Neuropharmacology* 2014; **80**: 115–132.
49 Lokk K, Modhukur V, Rajashekar B, Märtens K, Mägi R, Kolde R *et al.* DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol* 2014; **15**: r54.
50 Gillis J, Pavlidis P. The impact of multifunctional genes on 'Guilt by Association' analysis. *PLoS ONE* 2011; **6**: e17258.
51 R Studio, Inc. shiny: Easy web applications in R; 2014. Available at http://shiny.rstudio.com.
52 Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I *et al.* Recommendations for the design and analysis of epigenome-wide association studies. *Nat Methods* 2013; **10**: 949–955.
53 Walton E, Hass J, Liu J, Roffman JL, Bernardoni F, Roessner V *et al.* Correspondence of DNA methylation between blood and brain tissue and its application to schizophrenia research. *Schizophr Bull* 2015; **42**: 406–414.
54 Teh AL, Pan H, Chen L, Ong ML, Dogra S, Wong J *et al.* The effect of genotype and *in utero* environment on interindividual variation in neonate DNA methylomes. *Genome Res* 2014; **24**: 1064–1074.
55 McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S *et al.* Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol* 2014; **15**: R73.
56 Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, Gelfand ET *et al.* The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**: 648–660.
57 Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science* 2009; **324**: 929–930.
58 Li W, Liu M. Distribution of 5-hydroxymethylcytosine in different human tissues. *J Nucleic Acids* 2011; **2011**: e870726.
59 Lunnon K, Hannon E, Smith RG, Dempster E, Wong C, Burrage J *et al.* Variation in 5-hydroxymethylcytosine across human cortex and cerebellum. *Genome Biol* 2016; **17**: 27.
60 Lode Godderis CS. Global methylation and hydroxymethylation in DNA from blood and saliva in healthy volunteers. *BioMed Res Int* 2015; **2015**: 1–8.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017

Supplementary Information accompanies the paper on the *Translational Psychiatry* website (http://www.nature.com/tp)