Human Molecular Genetics, 2023, Vol. 32, 11, 1912–1921
https://doi.org/10.1093/hmg/ddad028
Advance access publication date 15 February 2023
Original Article

Rare genetic variants underlie outlying levels of DNA methylation and gene-expression

V. Kartik Chundru[1,2], Riccardo E. Marioni[3], James G. D. Prendergast[4], Tian Lin[1], Allan J. Beveridge[5], Nicholas G. Martin[6], Grant W. Montgomery[1], David A. Hume[7], Ian J. Deary[8], Peter M. Visscher[1], Naomi R. Wray[1,9] and Allan F. McRae[1,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia
[2]Wellcome Sanger Institute, Hinxton CB10 1RQ, UK
[3]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh EH4 2XU, UK
[4]The Roslin Institute, The University of Edinburgh, Midlothian EH25 9RG, UK
[5]Glasgow Polyomics, Wolfson Wohl Cancer Research Centre, The University of Glasgow, Glasgow G61 1QH, UK
[6]QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia
[7]Mater Research Institute, The University of Queensland, Brisbane, QLD 4102, Australia
[8]Lothian Birth Cohorts, Department of Psychology, The University of Edinburgh, Edinburgh EH8 9JZ, UK
[9]Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia

*To whom correspondence should be addressed at: Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia. Email: a.mcrae@uq.edu.au

Abstract

Testing the effect of rare variants on phenotypic variation is difficult due to the need for extremely large cohorts to identify associated variants given expected effect sizes. An alternative approach is to investigate the effect of rare genetic variants on DNA methylation (DNAm) as effect sizes are expected to be larger for molecular traits compared with complex traits. Here, we investigate DNAm in healthy ageing populations—the Lothian Birth Cohorts of 1921 and 1936—and identify both transient and stable outlying DNAm levels across the genome. We find an enrichment of rare genetic single nucleotide polymorphisms (SNPs) within 1 kb of DNAm sites in individuals with stable outlying DNAm, implying genetic control of this extreme variation. Using a family-based cohort, the Brisbane Systems Genetics Study, we observed increased sharing of DNAm outliers among more closely related individuals, consistent with these outliers being driven by rare genetic variation. We demonstrated that outlying DNAm levels have a functional consequence on gene expression levels, with extreme levels of DNAm being associated with gene expression levels toward the tails of the population distribution. This study demonstrates the role of rare SNPs in the phenotypic variation of DNAm and the effect of extreme levels of DNAm on gene expression.

Introduction

DNA methylation (DNAm) is involved in the regulation of gene expression (1–3). Variation in DNAm has been associated with many diseases, in particular cancers (4,5), but also common disease (6). Both genetic (7,8) and environmental (9–11) factors are highly influential to the variation in DNAm levels across the genome. In this study, we aim to characterize the effects of rare genetic single nucleotide polymorphisms (SNPs) on DNAm variation. This will help us in understanding the genetic architecture of DNAm, and potential mechanisms through which genetic variants can affect complex traits via effects on DNAm.

Variation in DNAm levels is known to be under partial genetic control; a family based study estimated the average heritability of DNAm levels to be $\underline{h^2} \sim 19\%$ (8), while another study estimated the average SNP-based heritability to be $h^2_{\underline{SNP}} \sim 21\%$ (12). DNAm quantitative trait loci (mQTL) analyses have discovered many associations between common genetic variants and DNAm levels across the genome (7,12–15). Regional control of DNAm has been observed in regions of up to 3 kb, through shared mQTL and correlations between DNAm levels across the region (7,16), while a Bayesian co-localisation study found evidence for a shared genetic effect between ~282 000 pairs of CpG-sites at a median distance of ~110 kb (14). Overlap between mQTL and gene expression QTL (eQTL) has also been observed (7,13), with genetic variants found to affect DNAm and gene expression levels pleiotropically (14,17). These observations point toward a possible mechanism through which genetic variants can alter gene expression levels via underlying differences in DNAm levels in a region.

Rare genetic variation has been shown to be important in the genetic architecture of complex traits, and gene expression (18–21). Rare variants have been found to be enriched near the transcription start site of genes in individuals with outlying levels of gene expression in both humans (22) and maize (23), and particularly in those individuals with outlying levels of gene expression across multiple tissue types (20), suggesting a large effect from rare genetic variation on gene expression levels. Some evidence has been found for similar effects of rare genetic variation near CpG-sites on DNAm variation (24), and in addition, using variant aggregation tests there has been evidence of effects from rare

Received: September 15, 2022. **Revised:** January 25, 2023. **Accepted:** February 9, 2023

variants on DNAm levels even when there is no common variant association at the relevant CpG-site (25). We hypothesize that, similar to the association found between rare variants and outlying gene expression levels (20,22,23), there are associations between rare variants and outlying levels of DNAm. Outliers in DNAm have been associated with common diseases such as motor neurone disease (26) and type I diabetes (27). Understanding the underlying mechanisms may help in determining the genetic aetiology of these associations between outlying DNAm and common disease. In addition, CpG-sites are known to be highly mutable, with the mutation rate at CpG-sites estimated to be one order of magnitude higher than anywhere else in the genome, which results in an enrichment of mutations at CpG-sites in the genome (28–34). Knowing how mutations at CpG-sites will affect DNAm and gene expression levels in the genome may also be important for understanding the genetic aetiology of complex trait diseases and cancers.

In this study, we aim to extensively examine the relationship between rare genetic variation and DNAm levels across the genome, and how DNAm levels may affect gene expression levels at nearby genes.

Results

An overview of the methods used in this study, with the different data available to us is given in Figure 1.

Detecting genome-wide genetic effects on DNAm

Using whole genome sequencing data and whole blood DNAm measures from the Illumina Infinium HumanMethylation450 array for $n = 1196$ individuals from the Lothian Birth Cohorts (LBC) of 1921 and 1936 (35), we tested for global effects of both rare and common genetic variants on DNAm levels across the genome. The number of minor alleles within 1 kb of the CpG-site were counted for each individual within a given minor allele frequency range, then, at each of the 415 007 DNAm probes, individuals were ranked from lowest DNAm level to the highest and number of minor allele counts was averaged at each rank across DNAm probes. As a control, before ranking the individuals at each DNAm probe, we randomly permuted the minor allele counts for each individual, which is equivalent to counting the minor alleles within a random 1 kb region in the genome. If there is no genetic effect on DNAm for SNPs with a given allele frequency range, we would expect no relationship between the average minor allele count across rank. We observe an inflation in allele counts at the lowest and highest ranks, for all minor allele frequency (MAF) ranges, while there is no inflation observed in the minor allele counts at a random 1 kb region (Fig. 2), suggesting genetic effects from variants in all MAF ranges affecting DNAm levels across the genome. This pattern of effect was also observed for minor alleles within 10 bp, 100 bp, 5 kb, 10 kb and 50 kb of the CpG-sites (Supplementary Material, Fig. S1).

For the common, and low-frequency variants ($MAF > 0.1$, and $0.01 < MAF < 0.1$, respectively), we show that these effects are largely captured by mQTL analyses by separating the ∼180 000 probes with a significant mQTL detected in previous studies (15), with no visible inflation at the ends of the distributions (Supplementary Material, Fig. S2). However, for the rare variants ($MAF < 0.001$, and $0.001 < MAF < 0.01$), the distributions after removing the mQTL probes remain inflated at the ends of the rank distribution, suggesting that mQTL do not capture the effects of rare variants.

The association between minor allele counts and DNAm rank is asymmetrical in Figure 2, with the lowest ranks having a larger inflation than the highest ranks in all MAF bins. This observation suggests a bias toward SNP minor alleles decreasing DNAm levels across the genome. However, after separating the probes which contain an SNP at the CpG-site (CpG-SNP) from the rest of the probes, we see that the inflations are more symmetrical for probes which do not contain a CpG-SNP, with slightly more inflation in the higher ranks (Supplementary Material, Fig. S3). This suggests that the allele disrupting the CpG site is, on average, the minor allele, which may be attributed to a combination of bias in selection of CpG sites included on the array (sites which are generally CpGs were chosen), and a known mutational bias in the genome from (methylated) cytosine to thymine through the process of deamination (33).

While inflation in the minor allele count is observed for individuals with either lowly or highly ranked methylation values for all MAF classes, for the rare variants ($MAF < 0.001$ and $0.001 < MAF < 0.01$), we see that the inflation is largely restricted to the extremes of the distribution. This is consistent with rare variants driving more extreme levels of DNAm.

Enrichment in rare alleles in individuals with outlying DNAm

We identified outlying DNAm levels at individual methylation probes using the subset of 613 individuals in the LBC dataset who have DNAm measurements at a minimum of three time-points. At a given time-point, an outlier was defined as a CpG-site in an individual with DNAm levels more than three times the interquartile range below the first quartile, or above the third quartile at that CpG-site. We detected a total of 3 698 676 outliers in at least a single time-point of measurement (each individual can be outlying at multiple probes). Approximately 80% (330 671/415007) of DNAm probes had at least one individual with outlying levels of DNAm. In addition, ∼5% of the outliers at a CpG-site (198 933/3698676) were consistently outlying at that site across at least three time-points. The outlier burden (mean number of outliers per individual at a time-point, Seeboth *et al.* 2019 (64)) was 2074 (out of 415 007 probes ∼0.5%), reducing to 304 (∼0.07%) when considering only those outliers stable across at least three time-points.

We observed an enrichment of ∼1.05× in the number of rare and low-frequency alleles within 1 kb of the CpG-site in outliers versus non-outliers. For the control analysis (randomly permuting the counts within 1 kb for each individual across all CpG-probes), no inflation was observed in any MAF range (Supplementary Material, Fig. S4).

We removed probes with a CpG-SNP as they may bias the enrichment. CpG-SNPs disrupt the methylation at the site which will likely result in outliers (36). We observe a much larger enrichment in all MAF groups when looking only at CpG-SNP probes (Supplementary Material, Fig. S5); however, we do not include these probes in any subsequent analyses.

The enrichment of rare alleles in outliers versus non-outliers stable across three to four time-points was larger relative to the transient outliers observed to be outlying at a single time-point (Fig. 3).

Those outliers which were stable, outlying at three to four time-points of measurement, had a much larger enrichment of minor alleles within 1 kb of the CpG-site than those outliers only observed to be outlying at a single time-point (Fig. 3).

The enrichment in the rare and low frequency alleles in stable outliers versus non-outliers was still significantly larger than 1,

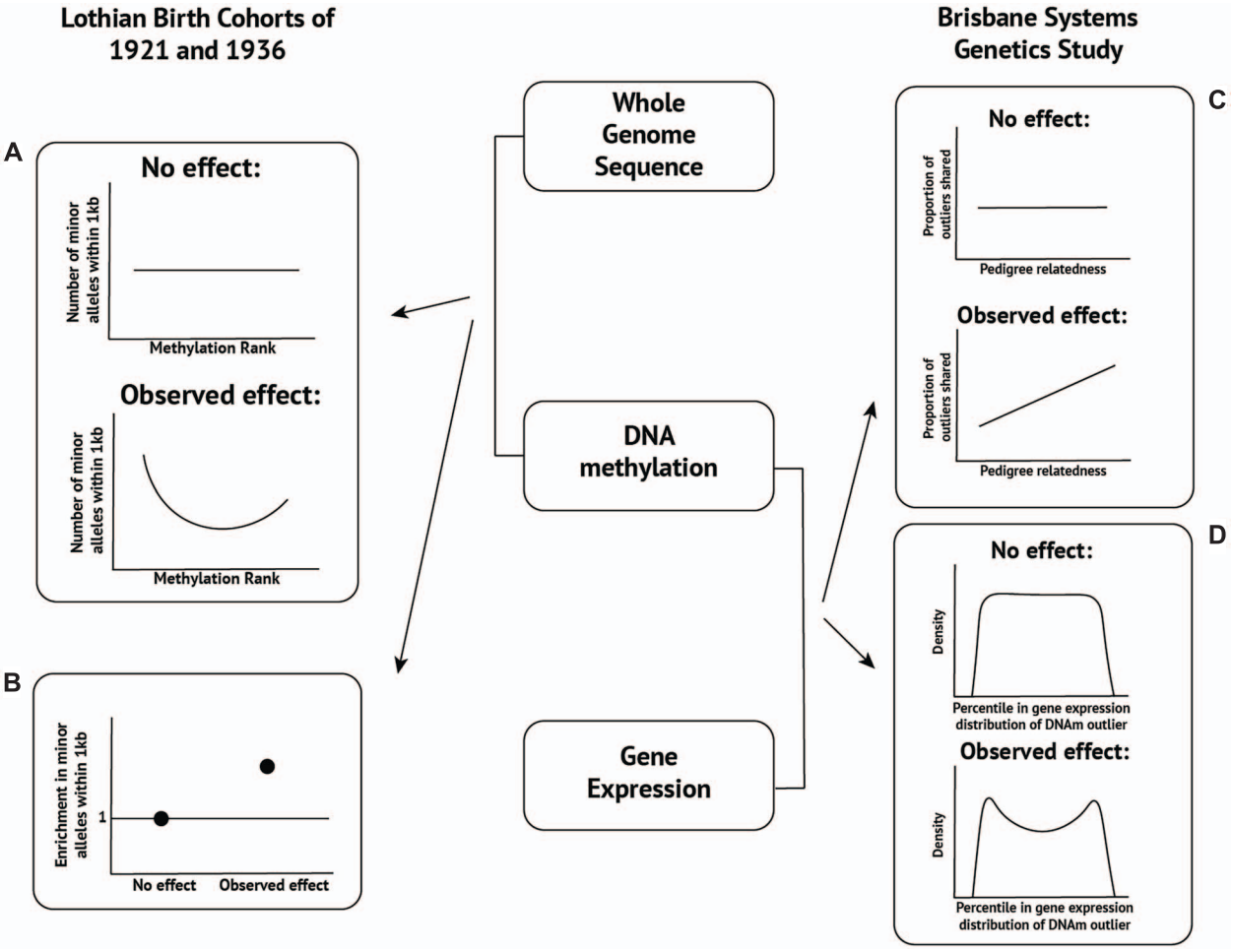

Figure 1. Overview of the methods used in this study. The Lothian Birth Cohorts of 1921 and 1936 were used to investigate the effect of genetic variants on DNA methylation levels, while the Brisbane Systems Genetics Study was used to examine the effect of DNA methylation levels on gene expression levels. (**A**) The number of minor alleles within 1 kb is plotted against methylation rank (The individual with the n^{th} lowest DNAm levels will have a methylation rank of n at that CpG-site); in the case of no effect of genetic variants on DNAm levels, a uniform distribution is expected, any deviation from the uniform distribution is evidence for a genetic effect on DNAm levels. (**B**) The enrichment of minor alleles within 1 kb in outliers compared to non-outliers; in the case of no effect of genetic variants on DNAm outliers, the enrichment will be 1, any significant deviation from 1 is evidence of an effect of genetic variants on outliers of DNAm. (**C**) The proportion of outliers shared between pairs is plotted against the pedigree relatedness; if there is no genetic effect on DNAm outliers a slope of 0 is expected, any non-zero slope is evidence for a genetic effect on DNAm outliers. (**D**) Finally, the distribution of gene expression percentile of individuals with DNAm outliers at nearby probes is plotted; in the case of no effect from DNAm on gene expression, a uniform distribution is expected, any deviation from the uniform distribution is evidence for an effect of DNAm on gene expression.

and similar from 10 to 50 kb with the enrichment not changing significantly (Supplementary Material, Fig. S6). The enrichment was much larger the closer we restrict from the CpG-site, although the confidence of the estimate is lower due to the smaller number of variants.

Outliers in gene expression and DNAm are shared between relatives

Using the Brisbane Systems Genetics Study (BSGS) dataset (37) ($n = 595$), which includes 67 MZ twin pairs, as well as many siblings and parent–offspring pairs with whole blood DNAm and gene expression array data, we detected a total of 1 133 080 outliers in DNAm levels (using the same definition of outliers as before), and 446 916 outliers in gene expression levels (using the definition of outliers as a gene expression probe in an individual with gene expression levels outside of 1.5× the interquartile range of the first or third quartile).

We observed a linear relationship between the proportion of DNAm outliers ($R^2 = 0.40$, slope = 0.18 and $P < 10^{-323}$) and gene expression outliers (adjusted $R^2 = 0.02$, slope = 0.03, $P < 10^{-323}$) shared between each pair of individuals, and their pedigree relatedness (Fig. 4). This is consistent with shared genetic effects underlying both outlying levels of DNAm levels and gene expression levels across the genome. However, there was very little overlap between gene expression outliers and DNAm outliers, with 6.1% of individuals with a gene expression outlier also having a DNAm outlier at the nearest annotated gene.

Outlying levels of DNAm are associated with a change in gene expression

Although the overlap of outlying DNAm and gene expression was not substantial, we tested whether the outlying DNAm levels correlated with any change in gene expression levels. For individuals with outlying levels of DNAm at a CpG-site, if the DNAm levels have no effect on gene expression levels, we would expect those individuals to be uniformly distributed across the gene expression distributions. Firstly, we paired DNAm probes to gene expression probes using significant common variant co-localization established using a summary data-based Mendelian randomization

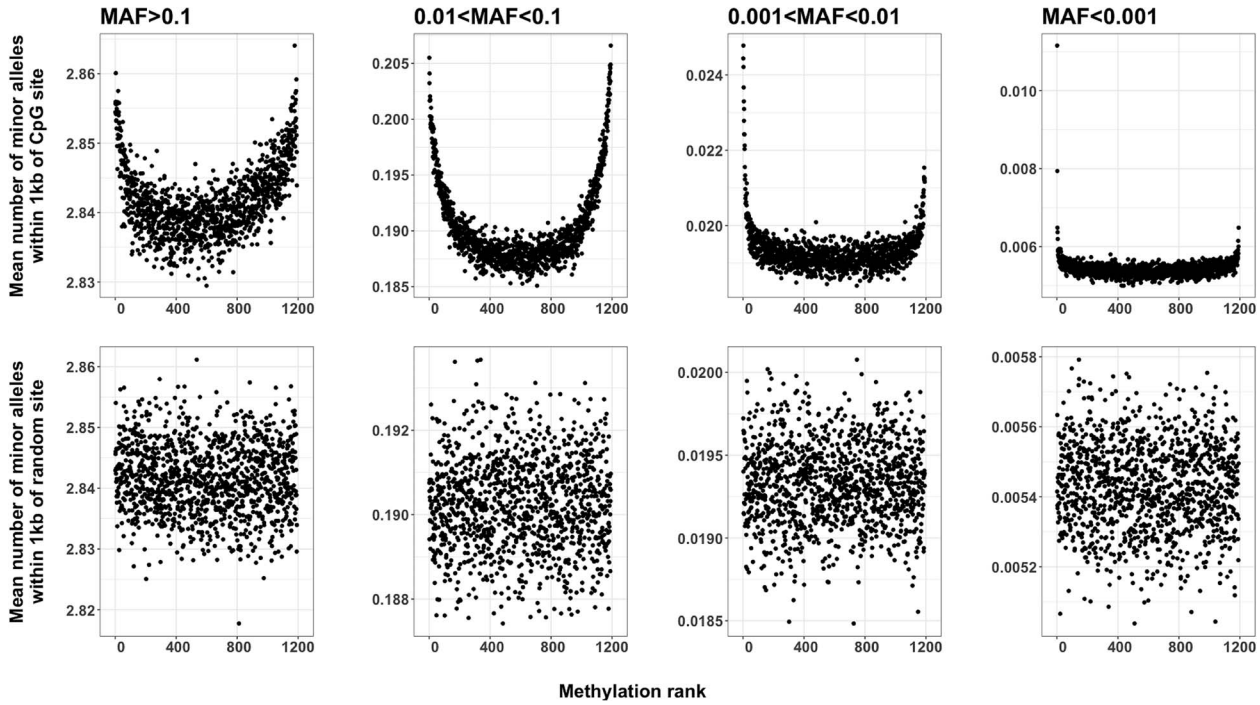

Figure 2. The mean number of minor alleles within 1 kb of the CpG-site, and a random site for each rank of DNAm levels across all autosomal probes across four MAF ranges. Each point represents the average number of minor alleles within 1 kb for the individual ranked xth from lowest to highest methylation levels for each of the 415 007 CpG-probes. Top: rare variants are associated with high and low ranking levels of DNA methylation. Bottom: no association remains after permutation of rare variant counts within individuals.

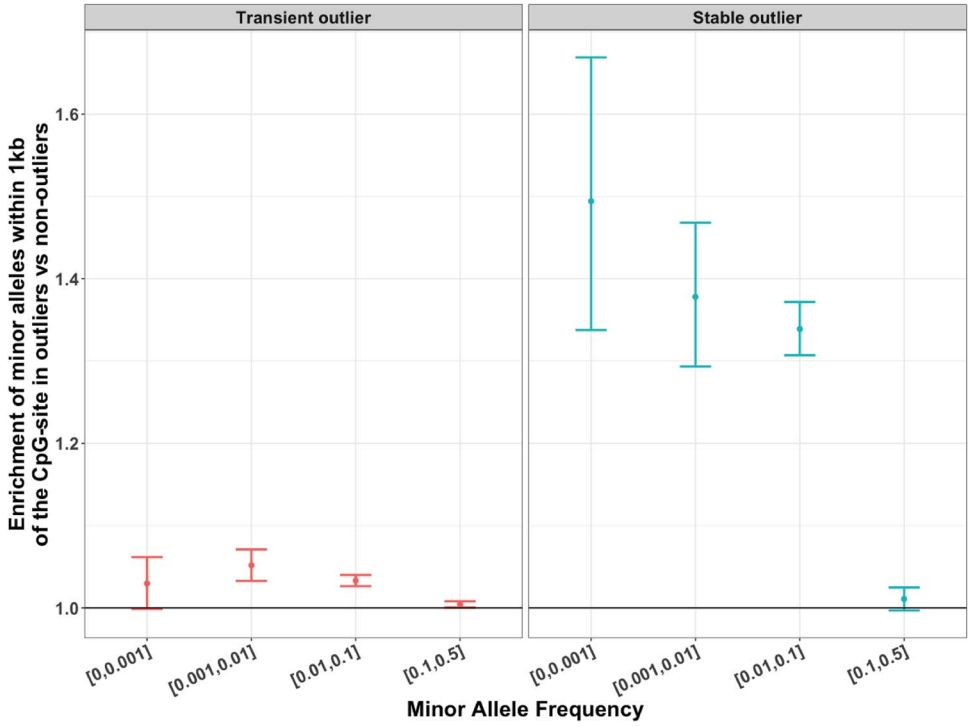

Figure 3. The enrichment in minor alleles within 1 kb of the CpG-site in transient and stable outliers vs non-outliers. The enrichment in rare alleles within 1 kb is significant for transient outliers in the rare ($0.001 < MAF < 0.01$) and low frequency ($0.01 < MAF < 0.1$) MAF bins, but substantially larger in outliers stable across time.

(SMR) study (17), then at DNAm probes within 10 kb of the gene expression probes. Wu *et al.* used SMR with DNAm as an exposure, and gene expression as an outcome in a Mendelian randomization framework to find a total of 10 588 associations between 7858 DNAm probes and 3239 gene expression probes after a stringent SMR *P*-value threshold and a relaxed HEIDI filter (heterogeneity in dependent instruments test, which filters out associations due to linkage). Using these linked probes allows us to focus on probe pairs already known to be linked by common variants.

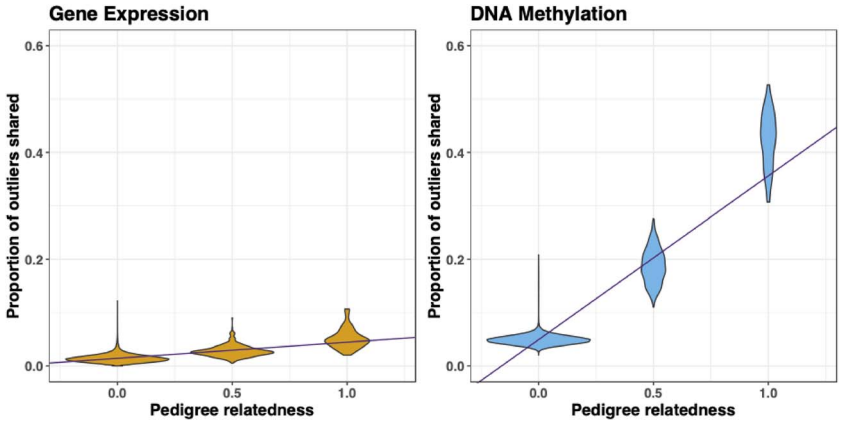

Figure 4. Outliers in DNAm and gene expression are shared between relatives more often than at random. The linear relationship between pedigree relatedness and proportion of outliers shared suggests a shared genetic component to the outlying levels of DNAm and gene expression. The difference in slope suggests a stronger genetic effect on the DNAm levels compared with gene expression levels.

The rank of gene expression levels for individuals with outlying methylation levels showed significant deviance from the uniform distribution with the individuals with outlying DNAm more likely to be at the ends of the gene expression distribution for both SMR-linked probes, and the probes within 10 kb (Cramér–von Mises test $\omega^2 = 6.48$, $P < 10^{-323}$, and $\omega^2 = 7.88$ and $P < 10^{-323}$, respectively, Fig. 5). The Cramér–von Mises test was nominally significant for the non-outlying individuals; however, the deviation from the uniform distribution was minimal and spread across the distribution, with the significance driven by the sample size. In contrast, for the outlying individuals, the deviation from the uniform distribution was in the tails of the distribution, and the magnitude of deviation was much larger than for the non-outlying individuals. The same pattern was observed at DNAm and gene expression probes within 1, 5, 10 and 50 kb of each other (Supplementary Material, Fig. S7). These results correspond to a correlation between outlying levels of DNAm and a change in gene expression levels at the relevant genes.

Stratifying the outliers by the direction, we see that those outliers which are higher than the sample mean correspond to a lower gene expression level, and the outliers lower than the sample mean correspond to a higher gene expression level. This shows the negative direction of effect from methylation to gene expression (Supplementary Material, Fig. S8).

Discussion

This study examined the links between DNAm levels, rare genetic SNPs and gene expression levels across the genome. We combined multiple lines of evidence to demonstrate the role of rare SNPs in outlying DNAm levels. Outlying levels of DNAm are further demonstrated to be associated with gene expression levels at nearby genes.

We examined the patterns of effects from common and rare genetic SNPs, within 10 bp, 100 bp, 5 kb, 10 kb and 50 kb of the CpG-site, on DNAm levels across the genome. We found that rare alleles were associated with extreme levels of DNAm. In addition, we observed a significant enrichment of rare alleles within 10 bp, 100 bp, 5 kb, 10 kb and 50 kb of CpG-sites in individuals with outlying levels of DNAm compared to individuals with normal DNAm levels at that CpG-site. Our results suggest that, in addition to common variants, rare variants also play a role in the control of DNAm levels across the genome.

DNAm levels at many CpG-sites are known to be correlated with age (16,38), and changes in environment are also known to have an effect across time (9–11). In our analysis, we found that outliers in DNAm levels which are present at only one time-point had almost no enrichment for rare alleles within 1 kb of the CpG-site compared to non-outliers, but those probes outlying across multiple time-points within an individual had significant enrichment, suggesting that transient outliers detected at a single time-point (3 177 418/3 698 676 ≈ 86% of the outliers in our study) are likely caused by environmental effects or measurement error, but the outliers stable across time are more likely to have an underlying genetic cause. This genetic effect underlying outliers in DNAm was confirmed using a family study design in an independent dataset. This is consistent with previous observations made using the LBC dataset in Shah *et al.* (39) who noted that many CpG-sites across the genome had stable DNAm across the lifetime, and these results are also in concordance with the observation made by Gaunt *et al.* (12) that the majority of mQTL are stable across time.

Similar to aggregation tests, we looked at enrichments and not associations with individual variants (which would be difficult to detect due to the power needed to reach statistical significance). We cannot say which variants have an effect and which do not. Only a single rare variant (MAF < 0.01) was observed within 1 kb of the CpG-site in over 19% (25 591/78484) of the outliers that were stable across time and had no CpG-SNP. However, even in these cases of only one rare allele within 1 kb, we cannot determine causality without functional experiments. In addition, our study only investigates SNP variants as this allowed for direct comparison with available mQTL studies. Structural variants have been associated with DNAm (e.g. 40,41) and potentially show larger effects than SNP variants. A further investigation of rare structural variants and their impacts of DNAm is warranted.

Previous studies have found correlations between DNAm and gene expression, and an overlap in the association of common genetic variants between them (7,13,42–46). In this study, we show that outliers in DNAm levels are associated with a difference in gene expression levels at nearby genes, with lower methylation levels corresponding to higher gene expression levels. Summary-data based Mendelian randomization (47) analyses have provided us with evidence of pleiotropic effects of common variants on DNAm and gene expression levels across the genome (14,17). In addition, the proportion of phenotypic variance explained by the lead variant at an mQTL was, on average, larger than the

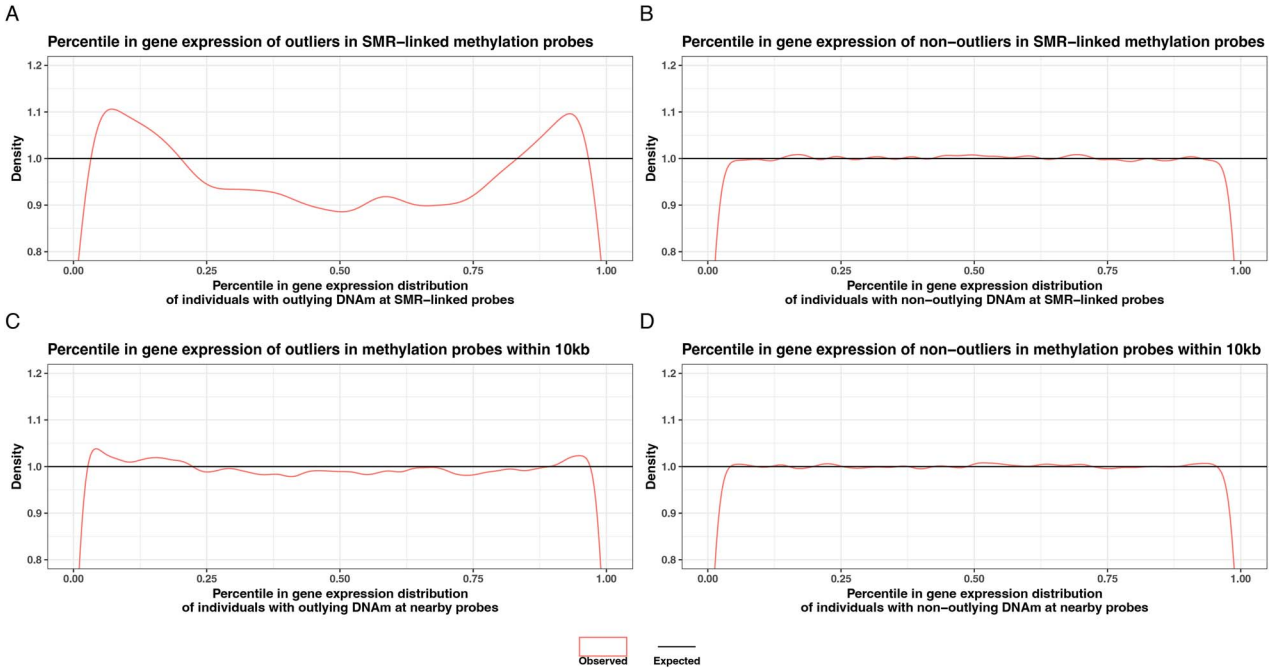

Figure 5. Density plot of the percentile in the gene expression distribution of individuals with outlying DNAm levels at linked DNAm probes, and at DNAm probes within 10 kb. Taking all DNAm and gene expression probe pairs linked through a SMR analysis (**A** and **B**), and at DNAm probes within 10 kb of a gene expression probe (**C** and **D**). There is significant deviation from the uniform distribution, with an inflation at the ends of the distribution, for gene-expression values in individuals with outlying levels of DNA methylation (A, C), and this deviation is most pronounced when gene and methylation sites are linked through SMR (A). There is minimal deviation from the uniform distribution of gene expression when considering individuals without outlying DNA methylation (B, D).

phenotypic variance explained by the same variant at a co-localized eQTL and at a co-localized higher-order complex trait QTL, such as height (17). This attenuation in effect size of the variant at each step suggests a mechanism of effect from genetic variant to DNAm, to gene expression, to higher-order complex trait. In this study, we observed that large differences in DNAm often corresponded to smaller differences in gene expression, which would fit into this hypothesized directional mechanism of effect. In addition, the difference in slope in Figure 4 also suggests a larger effect from genetic variation on DNAm levels than gene expression levels. However, given our study design of first identifying large effects on DNAm and then investigating their effects on gene expression, our effect size estimates will be biased by the winner's curse in DNAm, and care needs to be made in the interpretation of the direction of causation. Indeed, our observation is in contrast to some previous studies (e.g. 48) who arrive at the opposite conclusions when starting with discovery in eQTLs. More work will be needed to fully understand the interplay between DNAm and gene expression; however, it appears likely that DNAm plays both an active role in determining gene expression and a passive role in stabilizing it. The mechanism may be important to consider, as DNAm has been shown to be associated with many common diseases (6), and as methylation outliers are relatively easy to detect, it could provide a useful tool for future research.

A limitation of our study was that of the two data sets available to us, one (LBC) had WGS and DNAm array data, whereas the other (BSGS) had SNP array, DNAm array and gene expression array data. Ideally the study would be conducted on a cohort with all data types. With the increasing availability of whole genome sequence data, as well as RNA-seq and DNAm array/bisulphide sequence data, a more comprehensive study of the effects of rare variants on both DNAm and gene expression would provide a better understanding of the mechanisms underlying genetic effects on complex traits. We also transformed the DNAm and gene-expression data and corrected for various covariates. The rank-normal transformation helps limit false positive results, particularly when associating the measure with rare SNP variants. However, both the transformation and corrections for covariates affect the power of mQTL and eQTL detection and could lead to residual correlations between the measures. Limiting the scope of the analysis using SMR and sequence distance based filtering reduces the potential effects of this residual correlation on our results. Finally, this analysis was done using only white European individuals; while we do not expect the effect of rare genetic SNPs to differ across populations, there will be many more population specific effects compared to common variants.

Other epigenetic mechanisms, such as histone tail modifications, are highly correlated with DNAm levels, are under shared genetic control (7,13) and are also involved in the regulation of gene expression (46,49). We hypothesize that other epigenetic modifications may also show similar patterns of effects to what we found in DNAm, and including these into future analyses could potentially provide a more complete picture of the shared genetic control between DNAm, other epigenetic modifications and gene expression. A large proportion of patients who have genome sequencing undertaken are unsuccessful at being provided a molecular diagnosis (50,51). The improved functional annotation of noncoding variants is a particularly important step in identifying those variants that are truly pathogenic from those that are benign, resulting in improved diagnostic rates of disease genome sequencing studies while an understanding of the mechanism of effect. The combination of large sample sizes with extensive phenotypic records becoming available through many international biobanking efforts and high-throughput screening of genetic modification in cellular models will be needed to enable

the translation of the observation regional associations between rare genetic variants and outlying levels of DNAm through to determining individual rare variants and their mechanistic effect on disease.

In summary, this study provides a novel insight into the effect of rare variants on DNAm levels across the genome and shows that extreme differences in DNAm are associated with gene expression levels at nearby genes, which may be driven by rare genetic variation.

Materials and Methods

Lothian birth cohorts of 1921 and 1936

The LBC of 1921 and 1936 (LBC) (35) are longitudinal studies of cognitive ageing. DNA were extracted from whole blood samples from which DNAm levels were measured using the Illumina HumanMethylation450 BeadChip array across three or four time-points (the DNAm data was deposited in EGA with accession number EGAS00001000910). The raw intensity data were background corrected, corrected for cell-type and quantile normalized using standard QC protocols, and the DNAm beta-values were generated using the R package *meffil* (52). Probes were removed according to (53), removing probes with evidence of an overlapping SNP or repeat region, leaving 415 007 autosomal CpG-probes.

DNAm levels were measured at an average age (sd) of 79.1 (0.6), 86.7 (0.4) and 90.2 (0.1) years in the LBC1921 cohort and ages 69.6 (0.8), 72.5 (0.7), 76.3 (0.7) and 79.3 (0.6) years in the LBC1936. Of the 1342 individuals with DNAm measured at one point, 642 had at least three time-point measurements. While DNAm levels across the genome are known to change with age (16,38), this is not a confounding factor in our analysis as the age ranges within each wave of measurement are very narrow (mean standard deviation of age for each cohort in each wave was 0.6 years). Individuals with outlying B-cell counts were observed to have excess levels of outliers across all CpG-sites; these individuals were removed from the analyses.

Whole genome sequencing was performed on the HiSeq X with an average coverage of 36x (minimum 19.6×, maximum 65.9×) (The whole genome sequence data have been deposited in EGA with accession number EGAS00001003819). Details of the QC can be found in (54). Briefly, reads were mapped using BWA (55) to the build 38 of the reference genome, and GATK (56) was used for variant calling. Variant effect predictor (57) was used to annotate variants and gene models from the version 85 release of Ensembl.

To remove any distant relatives, we computed a genetic relatedness matrix from the common variants using GCTA (58), and removed one of every pair with value >0.05. Variants with VQSLOD<0.3546 were removed, and genotypes with GQ < 20 or DP < 7 were also removed. All samples are White European in ancestry, with no genetic outliers in the samples (Supplementary Material, Fig. S9).

The final sample size after taking individuals in both WGS and DNAm data after QC was $N = 1196$ and $N = 613$ for individuals with DNAm levels measured at three or more waves.

Brisbane Systems Genetics Study

The BSGS (37) was a dataset designed to study the genetic effects on gene expression, and the role of gene regulation in complex traits. DNAm levels were measured, in whole blood using the Illumina Infinium HumanMethylation450 BeadChip array (The DNAm data were deposited in NCBI GEO with accession number GSE56105), on 614 individuals from 117 families, including monozygotic twin pairs, dizygotic twin pairs, sibling pairs and parents. The QC of the DNAm data was performed using the same pipeline as with the LBC data. Gene expression levels were measured in whole blood on 846 individuals using the Illumina HumanHT-12 v4.0 BeadChip array (the gene expression data was deposited in NCBI GEO with accession number GSE53195). The QC of the gene expression data is detailed in (59). Briefly, the gene expression levels were normalized using variance stabilization (60), quantile normalized using the *limma* software (61), followed by PEER factor adjustment (62), with 50 factors, correcting for covariates such as age, sex, cell counts and batch effects. Both DNAm and gene expression levels were measured on a total of 595 individuals. All of these individuals are White European in genetic ancestry (37).

An overview of the methods used to investigate the effects of genetic SNPs on DNAm levels and gene expression levels using the LBC and BSGS datasets is shown in Figure 1.

Detecting genome-wide effects on DNAm

Following similar procedures to Zhao *et al*. (22) and Kremling *et al*. (23), we counted the number of minor alleles across all SNPs within 1 kb of the CpG-site in the LBC data for each individual, and we sorted the values by the rank of the individuals at each DNAm probe from lowest DNAm beta-value to the highest. We averaged this value at each rank across all autosomal probes to get the mean number of minor alleles within 1 kb of a CpG-site. We did this for four MAF ranges, MAF > 0.1, 0.1 > MAF > 0.01, 0.01 > MAF > 0.001 and 0.001 > MAF, which allowed us to separate the effects of common and rare variants. The rarest MAF bin (MAF < 0.001) corresponded to SNPs with one or two observed minor alleles in our dataset. This analysis was performed using the first wave of measurements in the LBC dataset to maximize sample size. The test was also repeated using larger distances from the CpG-site.

As a control analysis, before sorting by the rank at each DNAm probe, we randomly permuted the counts across CpG-sites for each individual which is roughly equivalent to counting the minor alleles at a random 2 kb region of the genome.

Detecting outliers

We defined DNAm outliers as a CpG-site in an individual with DNAm levels outside 3 interquartile ranges (IQRs) from the first quartile (Q1) or the third quartile (Q3) of the DNAm levels at that CpG-site, as in previous studies (63,64). The standard 1.5 IQRs from Q1 or Q3 compares to 3 standard deviations from the mean in a perfectly normal distribution. Our definition is slightly more stringent than this, as the distribution of DNAm levels can be highly skewed. For detecting outliers in the gene expression data, which had more symmetric distributions, the standard 1.5 IQR from Q1 and Q3 definition was used.

Enrichment of rare alleles around CpG-sites

We defined enrichment as

$$\text{Enrichment} = \frac{P\left(\text{individual is an outlier}\right)}{P\left(\text{individual is not an outlier at any timepoint}\right)}.$$

In words, we defined enrichment as the probability of an individual having a minor allele within 1 kb of a CpG-site given they have outlying DNAm levels at that site, divided by the probability of an individual having a minor allele within 1 kb of a CpG-site given they do not have outlying DNAm levels at that site. This is similar to the definition used in Li *et al*. (20), although they used a slightly

different definition of outliers (>2 standard deviations from the mean). We also repeated this test for larger distances from the CpG-site.

As a control analysis, considering only probes with at least one outlier, for each probe, we choose the outliers at a random probe to perform the count of individuals with a minor allele within 1 kb. This is equivalent to choosing a random 1 kb site in the genome but allows us to account for the CpG-sites on the DNAm array being more densely distributed in genic regions.

Proportion of outliers shared

To compute the proportion of outliers shared between each pair of individuals, we used the formula $\frac{2n_{12}}{n_1+n_2}$, where n_1 is the number of outliers for individual one, n_2 is the number of outliers for individual two and n_{12} is the number of outliers shared between the individuals. The relatedness coefficients were obtained from pedigree data.

Testing for association between outlying levels of DNAm and gene expression

To test for an association between outlying levels of DNAm and gene expression, the percentile in the gene expression levels distribution at a gene expression probe was calculated for each individual with outlying DNAm levels at the paired DNAm probe. We used two methods to pair DNAm probes to gene expression probes. First, we linked DNAm probes through a shared common variant co-localization with the gene expression probe detected using the SMR method (17,47). We also used all pairings of gene expression probes within 10 kb of the CpG-sites. This represents a trade-off between the number of pairs included in the analysis and including pairs of gene expression and DNAm probes that have no biological connection beyond proximity. Under the null hypothesis of no association between outlying DNAm and gene expression levels, the rank of gene expression levels for individuals with outlying DNAm levels should be uniformly distributed. We tested for deviation from the uniform distribution using the Cramér–von Mises test (65), which tests the degree of agreement between the sampled values and a theoretical distribution, in our case the uniform distribution.

As a control, we repeated this test with an equal number of non-outlying individuals at each probe.

Data access

The LBC methylation data used in this study have been submitted to the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) under accession number EGAS00001000910.

The LBC whole genome sequence data used in this study have been submitted to EGA (https://www.ebi.ac.uk/ega/home) under accession number EGAS00001003819.

The BSGS methylation data used in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE56105.

The BSGS gene expression data used in this study have been submitted to GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE53195.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We thank the cohort participants and team members who contributed to these studies. We would like to thank UQ RCC for their computing support. We would also like to thank L.A. Gough for her helpful comments on the manuscript.

Conflict of Interest statement. The authors declare that they have no competing interests.

Funding

UKs Biotechnology and Biological Sciences Research Council (BBSRC); The Royal Society; The Chief Scientist Office of the Scottish Government; Age UK (The Disconnected Mind project), Medical Research Council (grant MR/M01311/1); Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award); Age UK, The Wellcome Trust Institutional Strategic Support Fund; The University of Edinburgh; The University of Queensland; Medical Research Council and Biotechnology and Biological Sciences Research Council (grant MR/K026992/1 to I.J.D.); Australian National Health and Medical Research Council (NHMRC; grants 1010374, 1113400, 1010374, 496667, 1046880); Australian Research Council (ARC; grant DP160102400); NHMRC Fellowship Scheme (grants 1078037, 1078901 and 1083656 to P.M.V., N.R.W. and A.F.M.).

References

1. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
2. Fan, S. and Zhang, X. (2009) CpG island methylation pattern in different human tissues and its correlation with gene expression. *Biochem. Biophys. Res. Commun.*, **383**, 421–425.
3. Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33**, 245–254.
4. Feinberg, A.P., Koldobskiy, M.A. and Göndör, A. (2016) Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.*, **17**, 284–299.
5. Klutstein, M., Nejman, D., Greenfield, R. and Cedar, H. (2016) DNA methylation in cancer and aging. *Cancer Res.*, **76**, 3446–3450.
6. Jin, Z. and Liu, Y. (2018) DNA methylation in human diseases. *Genes Dis.*, **5**, 1–8.
7. Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K. and Gilad, Y. (2014) Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.*, **10**, e1004663.
8. McRae, A.F., Powell, J.E., Henders, A.K., Bowdler, L., Hemani, G., Shah, S., Painter, J.N., Martin, N.G., Visscher, P.M. and Montgomery, G.W. (2014) Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol.*, **15**, R73.
9. Christensen, B.C., Houseman, E.A., Marsit, C.J., Zheng, S., Wrensch, M.R., Wiemels, J.L., Nelson, H.H., Karagas, M.R., Padbury, J.F., Bueno, R. *et al.* (2009) Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.*, **5**, e1000602.
10. Dowen, R.H., Pelizzola, M., Schmitz, R.J., Lister, R., Dowen, J.M., Nery, J.R., Dixon, J.E. and Ecker, J.R. (2012) Widespread dynamic DNA methylation in response to biotic stress. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, E2183.
11. Garg, P., Joshi, R.S., Watson, C. and Sharp, A.J. (2018) A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PLoS Genet.*, **14**, e1007707.

12. Gaunt, T.R., Shihab, H.A., Hemani, G., Min, J.L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W.L., Ho, K. *et al.* (2016) Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.*, **17**, 61.
13. Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y. and Pritchard, J.K. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, **12**, R10.
14. Hannon, E., Gorrie-Stone, T.J., Smart, M.C., Burrage, J., Hughes, A., Bao, Y., Kumari, M., Schalkwyk, L.C. and Mill, J. (2018) Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *Am. J. Hum. Genet.*, **103**, 654–665.
15. McRae, A.F., Marioni, R.E., Shah, S., Yang, J., Powell, J.E., Harris, S.E., Gibson, J., Henders, A.K., Bowdler, L., Painter, J.N. *et al.* (2018) Identification of 55,000 replicated DNA methylation QTL. *Sci. Rep.*, **8**, 17605.
16. Bell, J.T., Tsai, P.-C., Yang, T.-P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A. *et al.* (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.*, **8**, e1002629.
17. Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., Lloyd-Jones, L.R., Marioni, R.E., Martin, N.G., Montgomery, G.W. *et al.* (2018) Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.*, **9**, 918.
18. Bomba, L., Walter, K. and Soranzo, N. (2017) The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.*, **18**, 77.
19. Hernandez, R.D., Uricchio, L.H., Hartman, K., Ye, C., Dahl, A. and Zaitlen, N. (2019) Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.*, **51**, 1349–1355.
20. Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J. *et al.* (2017) The impact of rare variation on gene expression across tissues. *Nature*, **550**, 239–243.
21. Marouli, E., Graff, M., Medina-Gomez, C., Lo, K.S., Wood, A.R., Kjaer, T.R., Fine, R.S., Lu, Y., Schurmann, C., Highland, H.M. *et al.* (2017) Rare and low-frequency coding variants alter human adult height. *Nature*, **542**, 186–190.
22. Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T.J., Lee, C.M., Banskota, S., Marigorta, U.M., Bao, G. and Gibson, G. (2016) A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.*, **98**, 299–309.
23. Kremling, K.A.G., Chen, S.Y., Su, M.H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J. and Buckler, E.S. (2018) Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, **555**, 520–523.
24. Barbosa, M., Joshi, R.S., Garg, P., Martin-Trujillo, A., Patel, N., Jadhav, B., Watson, C.T., Gibson, W., Chetnik, K., Tessereau, C. *et al.* (2018) Identification of rare de novo epigenetic variations in congenital disorders. *Nat. Commun.*, **9**, 2064.
25. Richardson, T.G., Shihab, H.A., Hemani, G., Zheng, J., Hannon, E., Mill, J., Carnero-Montoro, E., Bell, J.T., Lyttleton, O., McArdle, W.L. *et al.* (2016) Collapsed methylation quantitative trait loci analysis for low frequency and rare variants. *Hum. Mol. Genet.*, **25**, 4339–4349.
26. He, J., Tang, L., Benyamin, B., Shah, S., Hemani, G., Liu, R., Ye, S., Liu, X., Ma, Y., Zhang, H. *et al.* (2015) C9orf72 hexanucleotide repeat expansions in Chinese sporadic amyotrophic lateral sclerosis. *Neurobiol. Aging*, **36**, 2660.e2661–2660.e2668.
27. Paul, D.S., Teschendorff, A.E., Dang, M.A.N., Lowe, R., Hawa, M.I., Ecker, S., Beyan, H., Cunningham, S., Fouts, A.R., Ramelius, A. *et al.* (2016) Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat. Commun.*, **7**, 13555.
28. Aggarwala, V. and Voight, B.F. (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.*, **48**, 349–355.
29. Cooper, D.N. and Youssoufian, H. (1988) The CpG dinucleotide and human genetic disease. *Hum. Genet.*, **78**, 151–155.
30. Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C.M., Swertz, M., Wijmenga, C., van Ommen, G. *et al.* (2015) Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.*, **47**, 822–826.
31. Hodgkinson, A. and Eyre-Walker, A. (2011) Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.*, **12**, 756–766.
32. Moorjani, P., Amorim, C.E.G., Arndt, P.F. and Przeworski, M. (2016) Variation in the molecular clock of primates. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 10607–10612.
33. Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.
34. Ségurel, L., Wyman, M.J. and Przeworski, M. (2014) Determinants of mutation rate variation in the human germline. *Annu. Rev. Genom. Hum.*, **15**, 47–70.
35. Taylor, A.M., Pattie, A. and Deary, I.J. (2018) Cohort profile update: the Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.*, **47**, 1042–1042r.
36. Shoemaker, R., Deng, J., Wang, W. and Zhang, K. (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.*, **20**, 883–889.
37. Powell, J.E., Henders, A.K., McRae, A.F., Caracella, A., Smith, S., Wright, M.J., Whitfield, J.B., Dermitzakis, E.T., Martin, N.G., Visscher, P.M. and Montgomery, G.W. (2012) The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS One*, **7**, e35430.
38. Boks, M.P., Derks, E.M., Weisenberger, D.J., Strengman, E., Janson, E., Sommer, I.E., Kahn, R.S. and Ophoff, R.A. (2009) The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One*, **4**, e6767.
39. Shah, S., McRae, A.F., Marioni, R.E., Harris, S.E., Gibson, J., Henders, A.K., Redmond, P., Cox, S.R., Pattie, A., Corley, J. *et al.* (2014) Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.*, **24**, 1725–1733.
40. Shi, X., Radhakrishnan, S., Wen, J., Chen, J.Y., Chen, J., Lam, B.A., Mills, R.E., Stranger, B.E., Lee, C. and Setlur, S.R. (2020) Association of CNVs with methylation variation. *NPJ Genom. Med.*, **5**, 41.
41. Zhang, Y., Yang, L., Kucherlapati, M., Hadjipanayis, A., Pantazi, A., Bristow, C.A., Lee, E.A., Mahadeswar, H.S., Tang, J., Zhang, J. *et al.* (2019) Global impact of somatic structural variation on the DNA methylome of human cancers. *Genome Biol.*, **20**, 209.
42. Ball, M.P., Li, J.B., Gao, Y., Lee, J.-H., LeProust, E.M., Park, I.-H., Xie, B., Daley, G.Q. and Church, G.M. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.*, **27**, 361–368.
43. Deaton, A.M. and Bird, A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.
44. Lea, A.J., Vockley, C.M., Johnston, R.A., Del Carpio, C.A., Barreiro, L.B., Reddy, T.E. and Tung, J. (2018) Genome-wide quantification of the effects of DNA methylation on human gene regulation. *elife*, **7**, e37513.
45. Portela, A. and Esteller, M. (2010) Epigenetic modifications and human disease. *Nat. Biotechnol.*, **28**, 1057–1068.
46. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

47. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. and Yang, J. (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.
48. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S. *et al.* (2016) Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, **5**, e24.
49. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
50. Seaby, E.G. and Ennis, S. (2020) Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. *Brief. Funct. Genomics.*, **19**, 243–258.
51. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, Allen HL, Sanchis-Juan A, Frontini M Thys C *et al.* 2020. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature* **583**: 96–102.
52. Min, J.L., Hemani, G., Davey Smith, G., Relton, C. and Suderman, M. (2018) Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, **34**, 3983–3989.
53. Zhou, W., Laird, P.W. and Shen, H. (2016) Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.*, **45**, e22–e22.
54. Prendergast, J.G.D., Pugh, C., Harris, S.E., Hume, D.A., Deary, I.J. and Beveridge, A. (2019) Linked mutations at adjacent nucleotides have shaped human population differentiation and protein evolution. *Genome Biol. Evol.*, **11**, 759–775.
55. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
56. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
57. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. and Cunningham, F. (2016) The Ensembl Variant Effect Predictor. *Genome Biol.*, **17**, 122.
58. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
59. Lloyd-Jones, L.R., Holloway, A., McRae, A., Yang, J., Small, K., Zhao, J., Zeng, B., Bakshi, A., Metspalu, A., Dermitzakis, M. *et al.* (2017) The genetic architecture of gene expression in peripheral blood. *Am. J. Hum. Genet.*, **100**, 228–237.
60. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, S96–S104.
61. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
62. Stegle, O., Parts, L., Piipari, M., Winn, J. and Durbin, R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, **7**, 500–507.
63. Gentilini, D., Garagnani, P., Pisoni, S., Bacalini, M.G., Calzari, L., Mari, D., Vitale, G., Franceschi, C. and Di Blasio, A.M. (2015) Stochastic epigenetic mutations (DNA methylation) increase exponentially in human aging and correlate with X chromosome inactivation skewing in females. *Aging*, **7**, 568–578.
64. Seeboth, A., McCartney, D.L., Wang, Y., Hillary, R.F., Stevenson, A.J., Walker, R.M., Evans, K.L., McIntosh, A.M., Hägg, S., Deary, I.J. *et al.* (2020) DNA methylation outlier burden, health and ageing in generation Scotland and the Lothian Birth Cohorts of 1921 and 1936. *Clin. Epigenet.*, **12**, 49.
65. Arnold, T. and Emerson, J. (2011) Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal*, **3**, 34–39.