Identifying DNA methylation sites affecting drug response using electronic health record–derived GWAS summary statistics

Delaney A. Smith
Department of Biochemistry, Stanford University Medical School
Stanford, CA 94305, USA
Email: smithdel@stanford.edu

Stephanie A. Arteaga
Department of Biomedical Data Science, Stanford University Medical School
Stanford, CA 94305, USA
Email: sarteaga@stanford.edu

Marie C. Sadler
Department of Computational Biology, University of Lausanne
Lausanne, 1015, Switzerland
Email: marie.sadler@unil.ch

Russ B. Altman
Departments of Genetics and Bioengineering, Stanford University
Stanford, CA 94305, USA
Email: russ.altman@stanford.edu

Adverse drug responses (ADRs) result in over 7,000 deaths annually. Pharmacogenomic studies have shown that many ADRs are partially attributable to genetics. However, emerging data suggest that epigenetic mechanisms, such as DNA methylation (DNAm) also contribute to this variance. Understanding the impact of DNA methylation on drug response may minimize ADRs and improve the personalization of drug regimens. In this work, we identify DNA methylation sites that likely impact drug response phenotypes for anticoagulant and cardiometabolic drugs. We use instrumental variable analysis to integrate genome-wide association study (GWAS) summary statistics derived from electronic health records (EHRs) within the U.K. Biobank (UKBB) with methylation quantitative trait loci (mQTL) data from the Genetics of DNA Methylation Consortium (GoDMC). This approach allows us to achieve a robust sample size using the largest publicly available pharmacogenomic GWAS. For warfarin, we find 71 DNAm sites. Of those, 8 are near the gene *VKORC1* and 48 are on chromosome 6 near the human leukocyte antigen (*HLA*) gene family. We also find 2 warfarin DNAm sites near the genes *CYP2C9* and *CYP2C19*. For statins, we identify 17 DNAm sites. Eight are near the *APOB* gene, which encodes a carrier protein for low-density lipoprotein cholesterol (LDL-C). We find no novel significant epigenetic results for metformin.

Keywords: Pharmacogenomics; Pharmacoepigenetics, Biomarkers, DNA methylation, Electronic Health Records, Biobanks, Personalized Medicine.

1. Introduction

Adverse drug reactions (ADRs) lead to hundreds of thousands of deaths and hospitalizations each year.[1] Pharmacogenomic (PGx) studies show that genetic differences contribute to individual variance in response and are a source of ADRs because metabolic differences lead to higher-than-expected or lower-than-expected drug levels.[2] However, genetics alone do not explain all variance in drug response. Epigenetic modifications, such as DNA methylation (DNAm), have also been implicated.[3] For example, clopidogrel resistance is associated with DNA methylation near the genes BTG anti-proliferation factor 2 (*BTG2*), proteoglycan 2 (*PRG2*), vault RNA 2-1 (*VTRNA2-1*), and Period Circadian Regulator 3 (*PER3*).[4] While our DNAm profile may affect how we respond to many drugs, knowledge of specific interactions that allow prediction of variable drug response is limited.[3] Identifying methylation biomarkers for individual drugs may facilitate the reduction of adverse drug reactions.

PGx Genome-Wide Association Study (GWAS) reports have elucidated which genes and single nucleotide polymorphisms (SNPs) are associated with diverse drug response phenotypes.[5] However, these studies are limited by the fact that they do not account for epigenetic modifications. Pharmacoepigenetic (PEGx) studies, such as epigenome-wide association studies (EWAS) identify associations between DNAm and drug response phenotypes. However, these studies are limited both in number and statistical power. For example, there is currently one EWAS study on statins (linking statin use and type 2 diabetes, N = 6,820) in the EWAS catalog.[6,7] There are no studies on warfarin or metformin response.[7]

Instrument variable (IV) approaches are an alternative method to elucidate likely-causal interactions between an exposure (DNAm) and an outcome (drug response) from observational data.[8] Two sample methods allow researchers to integrate summary statistics from PGx GWAS studies with methylation quantitative trait (mQTL) data from separate sources to elucidate likely causal pharmacoepigenetic effects.[9] Moreover, analysis frameworks that use multiple IVs are less prone to reverse causality and artifacts arising from linkage disequilibrium (LD) patterns.[8] Mendelian randomization (MR) IV methods allow for the detection and elimination of pleiotropic markers while quantifying the direction and magnitude of causal effects (Figure 1).[8] This is key for PEGx studies because DNAm patterns change over time, making it challenging to dissect the cause, consequence, and confounding of PEGx effects.

This approach allows for -omics integration with existing PGx GWAS, identifying causal biomarkers such as DNAm. However, many existing PGx GWAS studies are underpowered (median sample size = 1220) for a robust statistical analysis.[10] While PGx GWAS statistics are more abundant than PGx EWAS reports, they still comprise only 10% of all GWAS entries in the GWAS catalog from 2016 to 2020.[10] A novel alternative method uses Electronic Health Record (EHR) data to generate GWAS summary statistics (Figure 1).[10] Biobank-generated summary statistics can have a large population size (UKBB N ~ 200,000) and have been shown to reflect PGx associations previously reported in traditional GWAS studies, albeit with weaker associations due to nosier phenotypes.[10]

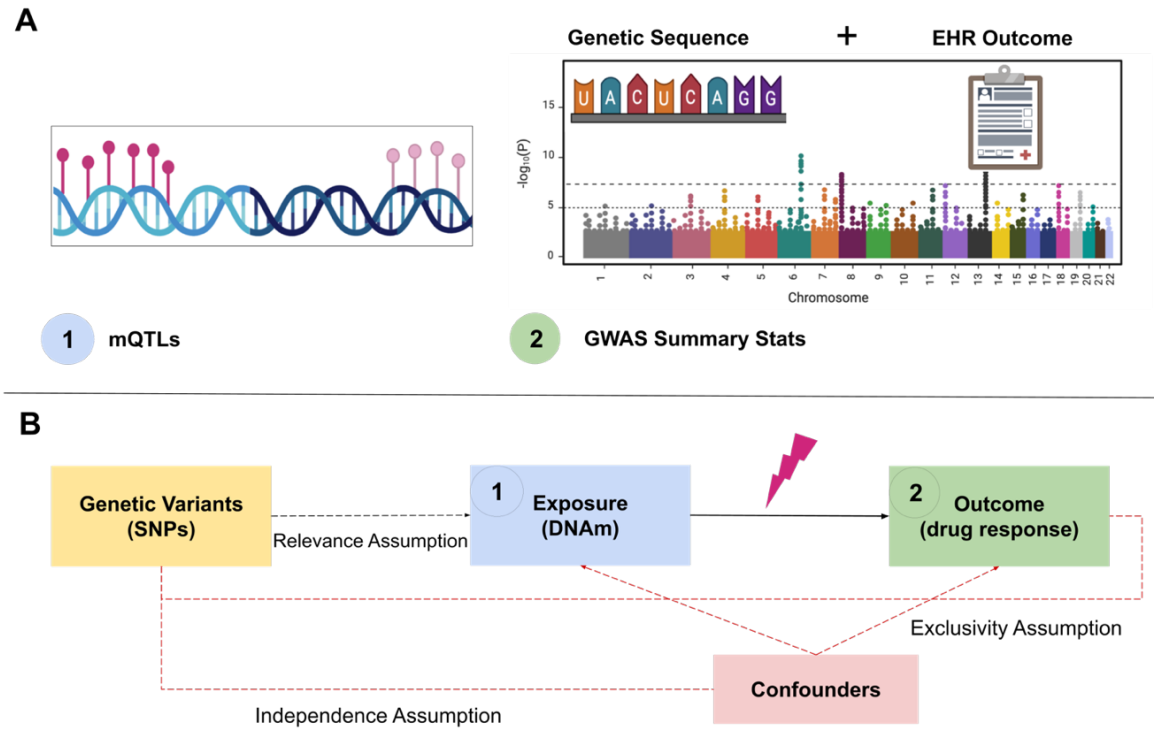

Fig. 1. Schematic of study design. A) mQTLs are taken from GoDMC in source 1 (left) and EHR records are combined with genetic sequences from the UKBB to generate summary statistics for data source 2 (right). B) Sources 1 and 2 are combined in a two-sample MR-IVW framework to determine the effect of DNAm on drug response (lightning bolt). The necessary assumptions are shown in dashed lines.

We demonstrate the efficacy of this approach in identifying DNAm sites that affect individual response to anticoagulant and cardiometabolic drugs. First, we analyze the effect of DNA methylation on warfarin response. Individual genetic differences of several genes, including vitamin K epoxidase reductase complex subunit 1 (*VKORC1*), cytochrome P450 family 2 subfamily C member 9 (*CYP2C9*), and member 19 (*CYP2C19*) are known to affect warfarin response.[11,12] Methylation near *VKORC1* has also been associated with differential warfarin response.[13] We also investigate the effect of DNA methylation on response to cardiometabolic drugs. Specifically, *β*–Hydroxy *β*–methylglutaryl-CoA (HMG-CoA) reductase inhibitors (common name: statins) and metformin. Individual response to these drugs is variable and is measured by low-density lipoprotein cholesterol (LDL-C) and total cholesterol (TC) levels for statins, and hemoglobin A1c (*HbA1c*) for metformin.[14,15] Some of this variance is explained by genetic factors such as variants in the apolipoprotein E (*APOE*) gene for statins and solute carrier family 2 member 2 (*SLC2A2*) for metformin.[16,17] In addition, metformin use is associated with genome-wide changes in DNAm levels, and a recent Swedish twin study revealed several DNAm sites associated with statin use.[18,19]

We report 69 total DNAm sites with an effect on warfarin response. Eight are near the gene *VKORC1*, and 2 are near *CYP2C19* and *CYP2C9*. Most (48) DNAm sites (also called CpGs) are not near known pharmacogenomic genes but are located on chromosome 6 near the *HLA* gene family. In the statin analysis, we find 8 CpGs near the apolipoprotein B (*APOB*) gene in addition to several CpGs near genes previously associated with cholesterol levels such as RING finger protein 39 (*RNF39*).[40,41] We find no novel significant epigenetic results for metformin. These findings allow us to better contextualize the role DNA methylation plays in individual drug responses.

2. Methods

2.1. *Genome-wide association summary statistics from electronic health records*

Genome-wide association summary statistics were generated from the EHRs of ~200,000 participants of the UKBB,[20] as described in Sadler *et al.* 2024.[10] Briefly, longitudinal medication patterns were analyzed to identify drug type, dose regimens, and drug adherence as well as baseline and post-treatment biomarker levels. We used the following pharmacogenetic phenotypes: average warfarin daily dose over the past five prescriptions (N = 4,554; McInnes and Altman), [21] cholesterol response to statins (N = 26,669 for TC, N = 17,063 for LDL-C),[10] and *HbA1c* response to metformin (N = 4,119).[10] GWAS on these quantitative traits were conducted with the REGENIE software (v3.2.4) in a whole-genome regression model for genetic markers with a minor allele frequency (MAF) > 0.05.[22] SNPs in high LD regions were removed along with those not passing LD pruning at $r^2 < 0.9$.[8]

2.2. *Two sample summary statistic instrument variable analysis*

We conducted two sample summary statistic instrument variable analyses using an inverse variance-weighted framework.[8] We used the SMR-IVW software (v1.0) as it allows two-sample IVW analysis with GWAS summary statistics.[8] The settings were: p-value (p) of mQTLs < 1×10^{-6}, LD $r^2 < 0.01$, *cis* window range of 1,000 kilobases (Kb), and the LD matrix was included in causal effect calculations. The tolerated allele frequency difference for each SNP between datasets was 0.1. A Steiger filter was implemented as described in Hemani *et al.* 2017 with a threshold set at -2, equivalent to a one-sided t-test p-value threshold of 0.023.[23] This strict threshold diminished the likelihood of including reverse causal relationships. We used mQTL data from the GoDMC database (N = 32,851), which contains > 170,000 whole blood DNAm sites with at least one significant *cis*-mQTL ($p < 1\times10^{-6}$, < 1 Mb from the DNAm site, N > 5,000).[24] The LD reference panel was from the 1,000 Genomes Project.[25]

2.3. *Multiple hypothesis correction*

To correct for multiple hypothesis testing, we used a false discovery rate ($\alpha = 0.05$) calculated by the Benjamini-Hochberg method from the statsmodels.stats.multitest (v0.14.2) package for Python.[26]

2.4. *Sensitivity to pleiotropy and heterogeneity analyses*

All CpGs that passed the significant threshold were also pruned to ensure a minimum of 3 instrumental variables. The remaining CpGs underwent sensitivity analysis. We first calculated a Cochran's Q statistic using the Metagen R package (v4.9.6),[27] and the corresponding Chi-Squared distribution p-value using the R Stats Chi-Square function (v3.6.2).[28] We removed any CpGs with significant evidence of heterogeneity ($p < 0.05$). [8] Next, we calculated an F-statistic (F) and removed any results with evidence of weak instrument bias ($F < 10$). [8] We tested for evidence of horizontal pleiotropy by analyzing the intercept values of an MR-Egger regression using the 2SMR package (v0.6.6).[23] Any CpGs showing significant evidence of pleiotropy ($p < 0.05$) were removed.

Table 1. Results of warfarin GWAS integration. For brevity, CpGs are displayed together if they are within approximately 1 Mb. When multiple CpGs are grouped, the CpG information represents the signal with the highest absolute value effect size. Full results are available on GitHub: https://github.com/smithdelaney/PGx-MR-from-EHR-GWAS.

CpG	Location	Number of CpGs	β	*SE*
cg06617202	1: 205038787	2	0.289	0.071
cg03935872	2: 17935919	1	0.349	0.086
cg06197503	3: 36422406	1	-0.286	0.063
cg10961486	4: 69959004	1	0.145	0.035
cg27585641	5: 73024506	2	0.336	0.077
cg15601071	6: 30078080	48	-0.768	0.183
cg13455759	9: 119655874	1	0.128	0.032
cg07530925	10: 90564681	1	-0.200	0.050
cg15404570	10: 96943130	1	0.597	0.127
cg03708694	11: 44489577	1	0.407	0.093
cg05555928	11: 63887634	1	0.110	0.027
cg08374890	16: 31117067	8	0.799	0.030
cg04077706	19: 43442484	1	0.128	0.030

3. Results

3.1 Warfarin MR-IVW results

Genome-wide hypothesis correction revealed 76 CpGs which exceeded the significance threshold. Of these, 69 CpGs showed no evidence of pleiotropy, weak instrument bias, or heterogeneity and were considered for further analysis (Figure 2A, Table 1). Eight of these CpGs were *cis* (within 0.5 megabases (Mb)) to the gene *VKORC1* (Figure 2A). Four had a positive beta (β) value (causing a higher warfarin dose) and 4 had a negative β (causing a lower warfarin dose), with absolute value effect sizes ranging from $|\beta| = 0.314$ to $|\beta| = 0.799$. The average absolute-value effect size was $\overline{|\beta|} = 0.554$ with an average standard error ($\overline{SE}$) of 0.046. One CpG (cg15404570) was *cis* to *CYP2C9* and *CYP2C19* and had a positive effect size $\beta = 0.597$ and $SE = 0.127$. Forty-eight CpGs (70% of all significant CpGs) were located on the short arm of chromosome 6, between 28.3 and 31.1 Mb (Figure 2B). These signals are *cis* to genes encoding the tripartite motif (*TRIM*) protein family and the HLA protein family (Figure 2B). Twenty-seven of these CpGs had a positive effect size and 21

had a negative effect size ($|\bar{\beta}| = 0.204$, $\overline{SE}) = 0.0475$). Absolute value effect sizes ranged from $|\beta| = 0.057$ to $|\beta| = 0.768$.

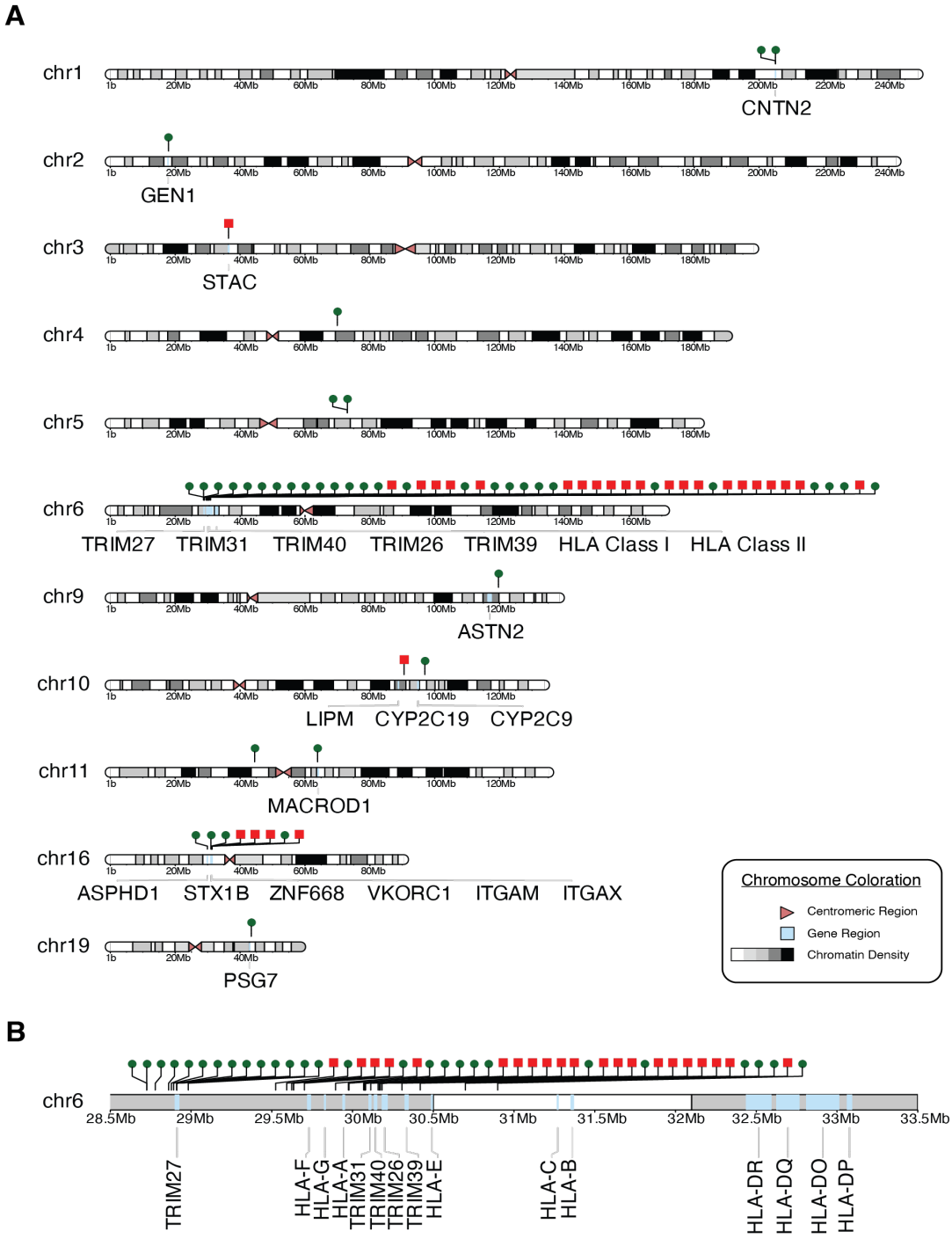

Fig. 2. A) Display of 69 CpGs found in the warfarin analysis. CpGs with a positive effect, or higher warfarin dose, ($\beta > 0$) are shown in green, and those with a negative effect ($\beta < 0$) are shown in red. Nearby genes are also annotated. B) Zoomed-in representation of the short arm of chromosome 6.

3.2 Statin MR-IVW results

GWAS integration results revealed 8 significant CpGs with LDL-C reduction as the outcome and 10 additional CpGs for TC reduction. Following quality control testing, 17 CpGs were further analyzed (Figure 3, Table 2). The 8 CpGs derived from the LDL-C analysis (47% of all CpGs) were *cis* to the *APOB* gene, which encodes an LDL-C carrier protein. All effect sizes for these 8 CpGs were negative, with an average absolute value of $|\overline{\beta}| = 0.088$, and an average standard error of $\overline{SE} = 0.018$. In this case, a negative β means that statin efficacy is increased since the clinical goal of the therapeutic is to reduce cholesterol levels. The absolute value effect size ranged from $|\beta| = 0.079$ to $|\beta| = 0.108$. Two CpGs (cg05337441, cg24309555) were previously annotated for TC or LDL-C in the EWAS catalog.[7] There were 3 CpGs (cg06028875, cg16908633, cg23752348) on the short arm of chromosome 6 which were *cis* to the *RNF39* gene and near the *HLA* gene family (within 1 Mb) (Figure 3). All 3 had a negative effect ($|\overline{\beta}| = 0.075$, $\overline{SE} = 0.016$). Five CpGs were on chromosome 10, four of which had a positive effect size ($|\overline{\beta}| = 0.059$, $\overline{SE} = 0.015$). The magnitude of these effect sizes ranged from $|\beta| = 0.031$ to $|\beta| = 0.158$. Additional genes associated with TC CpGs were DPY30 domain-containing proteins 1 and 2 (*DYDC1/C2*), erythroblast transformation-specific (ETS) proto-oncogene 2 (*ETS2*), tetraspanin 14 (*TSPAN14*), and peroxiredoxin-like 2A (*PRXL2A*) (Figure 3).

Table 2. Results of statin GWAS integration. For brevity, CpGs are displayed together if they are within approximately 1 Mb of each other. When multiple CpGs are grouped, the β is the absolute value average and the CpG name and location represent the signal with the highest absolute value effect size. Full results are available on GitHub: https://github.com/smithdelaney/PGx-MR-from-EHR-GWAS.

CpG	Location	Number of CpGs	β	*SE*
cg00673290	2:21266727	8	-0.108	0.020
cg06028875	6: 30042295	3	-0.087	0.019
cg02750471	10: 82179740	5	0.158	0.031
cg15892280	21:40180000	1	0.088	0.019

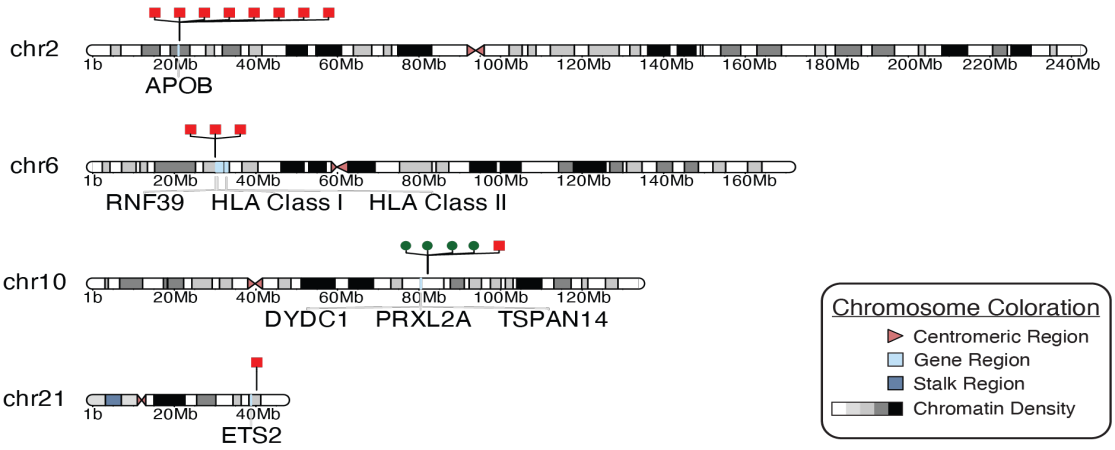

Fig. 3. Display of 17 CpGs found in the statin analysis. CpGs with a positive effect ($\beta > 0$) are shown in green and those with a negative effect ($\beta < 0$) are shown in red. A negative effect in this study means increased statin efficacy. Nearby genes are also annotated.

Discussion

In this study, we provide evidence that DNA methylation plays a causal role in individual response to warfarin and statins. Probing PEGx effects using EWAS studies provides correlative associations between DNAm sites and drug response phenotypes. Our approach uses existing information to infer directional, causal, and quantitative effect estimates. In our warfarin analysis, 8 CpGs were *cis* to *VKORC1*. Warfarin's mechanism of action targets *VKORC1*, and genetic variations in *VKORC1* are known to modulate warfarin's efficacy.[11,12] Recent findings also implicate *cis* DNA methylation near *VKORC1* in warfarin response.[13] In addition, we find 8 CpGs *cis* to *APOB* in the statin analysis. *APOB* encodes an LDL-C carrier and has 187 GWAS associations in the GWAS Catalog with LDL-C and 125 with TC.[29] Two of the CpGs we identified also had previous annotations for LDL-C or TC.[7] Thus, our approach captures the known effects of DNA methylation on drug response. We also show that in the case of metformin, no novel significant DNA methylation effects were detected. This could be because there is no biological effect or because the GWAS is underpowered.[10]

Our method depends on the assumptions underlying IV analysis. The first assumption is that there is a sufficiently strong relationship between the instrumental variables (SNPs) and the exposure (DNAm). The second assumption (independence) is that instrumental variables are independent of confounders. The third assumption (exclusivity) is that any effect the SNP has on the outcome is mediated only through the exposure (no horizontal pleiotropy). The first assumption can be tested by selecting highly significant mQTL effects ($p < 1\times10^{-6}$) and performing a weak instrument bias test (F-statistic).[8] The second and third assumptions are violated when results show evidence of heterogeneity, horizontal pleiotropy, or the presence of invalid instruments. We use Cochran's Q test to detect evidence of heterogeneity, and the presence of invalid instruments.[30] MR-Egger regression intercepts detect the presence of horizontal pleiotropy.[9] In addition, these assumptions may not hold in the presence of LD between the mQTLs and SNPs. The risk of 'LD-hitchhiking' leading to spurious results is managed by selecting CpGs with a minimum of 3 instrumental variables, filtering

out SNPs in high-LD regions, pruning for independence, and using a Steiger filter for directionality.[8] In our analysis, we excluded 7 warfarin CpGs and 1 statin CpG which did not pass all these controls. However, the possibility of horizontal pleiotropy can never be fully excluded.

Our results show DNA methylation CpGs *cis* to warfarin pharmacogenomic genes, *VKORC1*, *CYP2C9*, and *CYP2C19*.[11,12] Therefore, individuals' methylation profiles may account for some of the variability in warfarin response not captured in pharmacogenomic models. The 8 *VKORC1* CpGs had effect sizes in both positive and negative directions, indicating that while the presence of some CpGs reduced the average daily dose of the patient, others likely led to an increase. Seventy percent (70%) of effect CpGs were located near the *HLA* and *TRIM* genes on chromosome 6 (Figure 2B). The CpGs were closest to several *TRIM* genes (*TRIM26, TRIM27, TRIM31,* and *TRIM40).* These genes encode proteins that have varied and widespread functionality. DNAm may regulate the expression of *TRIM* genes, which have many downstream effects, possibly including modulating blood clotting pathways. However, recent work on DNA methylation and gene expression shows that methylated sites can act distally to influence the expression of neighboring genes.[3,24,31,32] Thus, an alternative hypothesis is that these CpGs impact the expression of the *HLA* genes, which are interlaced with *TRIM* genes on chromosome 6 (Figure 2B). Genetic polymorphisms in *HLA* genes have previously been associated with blood disorders including acquired hemophilia A, venous thrombosis, immune thrombotic thrombocytopenic purpura, and sickle cell disease.[33-36] Therefore, the pre-existing association between *HLA* genes and blood diseases may manifest through altered warfarin response, which is affected by DNA methylation near these genes.

Our statin results show that causal CpG identified in this study are not located near known pharmacogenomic genes. Instead, 47% of CpGs are *cis* to *APOB,* which encodes an LDL-C carrier (Figure 3). These CpGs all have an average effect size of $|\bar{\beta}|$ 0.087, all with a negative direction, meaning that methylation in this region causes a decrease in measured LDL-C in response to starting statin treatment. Therefore, the presence of these CpGs causes improved efficacy of statin treatment.

Another apolipoprotein gene, *APOE*, has over 20 variant annotations for statin efficacy in the PharmGKB pharmacogenomic database.[37] Both *APOB* and *APOE* are carriers of LDL-C and other lipoproteins. While *APOB* does not have PharmGKB annotations for statins, there are several genetic variants within the gene associated with LDL-C and TC levels in the GWAS Catalog, as discussed above. Moreover, genetic variation in *APOB* has been associated with familial hypercholesterolemia,[38] and levels of *APOB* are biomarkers of atherogenic particle concentration in the bloodstream (Figure 4).[39] Since DNAm near the *APOB* gene causes decreases in LDL-C in response to statin treatment, these CpGs can be biomarkers of statin response, and the study of these CpGs can increase our understanding of atherogenic disease.

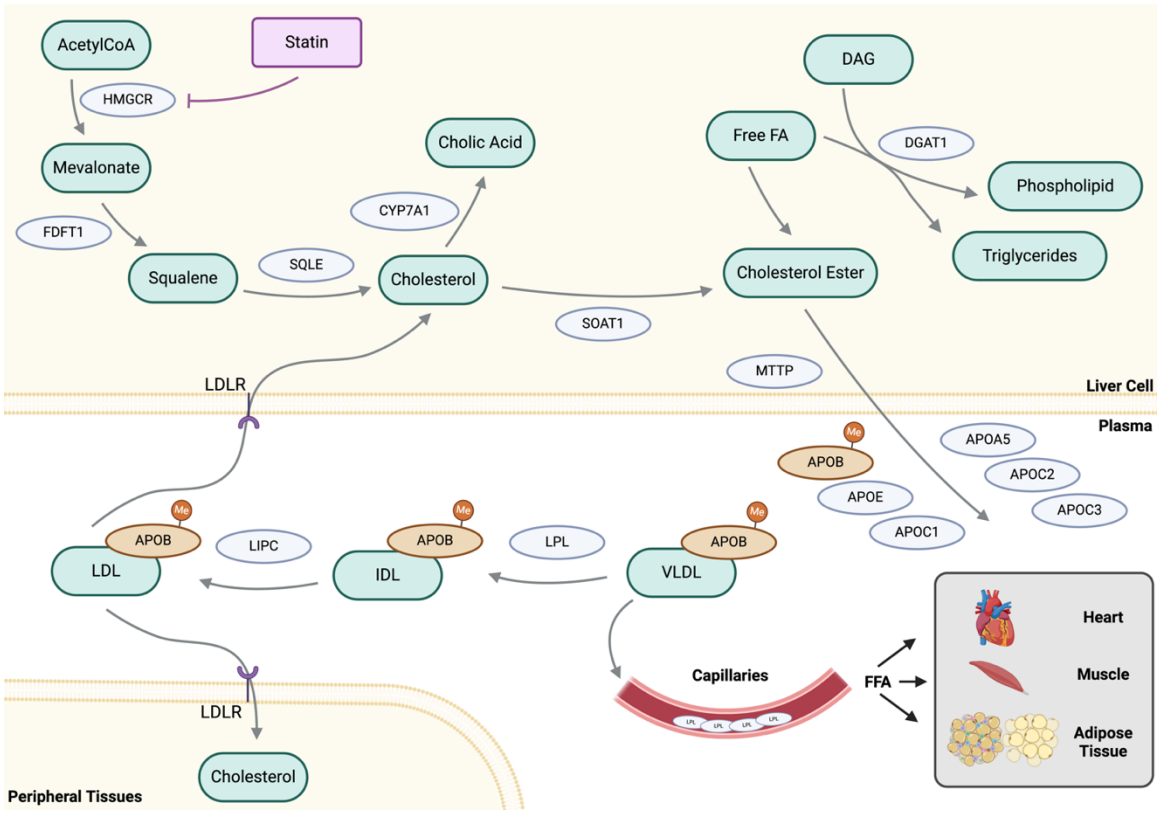

Fig. 4. The statin pharmacodynamic pathway and *APOB*-LDL-C pathway. Purple indicates statin, small molecules are in green, and light blue shapes represent protein-coding genes with HGNC standard names. Orange 'Me' probes represent DNA methylation occurring near the gene coding region (created with BioRender.com; adapted from PharmGKB[37]).

We propose a potential model of how DNAm near *APOB* may affect LDL-C levels in response to statin therapy (Figure 5). The 8 CpGs we identified near *APOB* are in the gene regulatory region. Five CpGs (cg16306978, cg16723488, cg24309555, cg25071744, cg25123895) are in the *APOB* promoter region, one in an enhancer region (cg05337441), and one (cg00673290) in a CpG island within the regulatory region. Generally, DNAm within the regulatory region of a protein-coding gene is associated with decreased expression.[54] Reduced *APOB* can lead to an increased ratio of LDL-C to *APOB*.[55-56] With statin therapy, LDL receptor expression increases as intracellular hepatic cholesterol decreases (Figure 5). Since *APOB* binds to the LDL receptor, more LDL-C is cleared from the plasma per *APOB* particle, leading to a greater decrease in measured LDL-C (Figure 5). It could also be that an individual with reduced *APOB* levels stores more LDL-C in other cholesterol-carrying particles. If these particles are equally reduced with statin therapy, then folks with higher *APOB* levels will have higher post-treatment LDL-C levels. Moreover, cholesterol metabolism is an intricate pathway, and regulatory mechanisms are still being studied, so additional experiments would be required to test these hypotheses.

The statin results also show that the gene *RNF39* had 3 nearby CpG sites. *RNF39* is involved in inflammatory responses throughout the body and has a SNP that has previously been associated with free cholesterol levels.[40,41] We find 2 CpGs near *DYDC1/C2*. These genes are primarily studied for their role in spermiogenesis, but several SNPs in the gene have been previously associated with hypertension.[42,43] One CpG was near the *TSPAN14* gene, which is associated with Niemann-Pick disease, a genetic disorder that leads to the inability to break down fats, such as cholesterol and lipids, inside cells.[14] Another CpG was near *ETS2*, which is a transcription factor. It regulates the transcription of proteasome assembly chaperone 1 (*PSMG1*) which has two SNPs associated with LDL in the GWAS catalog.[7] Finally, the CpG with the largest absolute effect size ($|\beta| = 0.158$) was located near gene *PRXL2A*. This gene interacts with ST3 beta-galactoside alpha-2,3-sialyltransferase (*ST3GLA4*) which has 48 SNPs associated with LDL-C and 32 SNPs associated with TC in the GWAS catalog.[29] These associations provide plausible pathways by which DNA methylation may impact response to statin treatment.

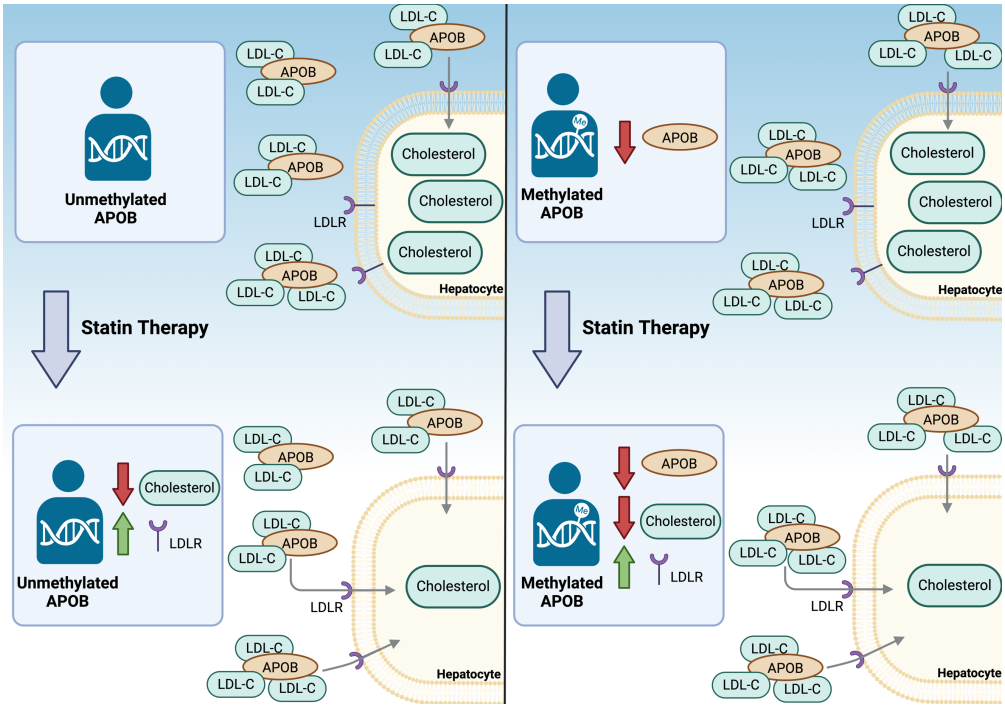

Fig. 5. Lower *APOB* expression may lead to decreased LDL-C levels after statin therapy due to an increased ratio of LDL-C to *APOB* (created with BioRender.com).

We observed that 5 of the LDL-C CpGs (cg16306978, cg24309555, cg25035485, cg25071744, cg25123895) and 3 TC CpGs (cg01528321, cg02750471, cg04043334) identified in the statin analysis (8 total, 47%) had previous annotations for inflammatory disease (inflammatory bowel syndrome (IBD) and Crohn's disease) in the EWAS catalog.[44] A comprehensive EWAS study has published approximately 3,633 CpGs associated with either disease, which make up about 2.1% of all CpGs in GoDMC.[44] The number of overlapping annotations is significantly ($p < 1\text{x } 10^{-8}$) greater than what is expected due to random chance alone. These findings, in combination with the warfarin

CpGs located near *HLA* genes, suggest that there are shared pathways between immune response and response to common cardiovascular and clotting disorder treatments that may be influenced by DNAm patterns. This mirrors other recent findings that are beginning to dissect how these two systems interact outside of the epigenetic space.[45-47] However, since the *HLA* gene region has high genetic diversity, it is possible that the signals detected reflect differences in ancestry and prevalence of *HLA* haplotypes. Thus, we also examined the overlap between the statin DNAm sites, and those annotated for inflammatory diseases, excluding the sites in the *HLA* region (non-*HLA* sites: cg16306978, cg24309555, cg25035485, cg25071744, cg25123895) and found that the overlap remained significant ($p = 0.052$).

This study had several limitations. First, we analyze whole-blood DNA methylomes. DNA methylation is tissue-specific and much of the pharmacokinetic and pharmacodynamic activity occurs in the liver. While some genes have similar DNA methylation patterns across blood and liver, this assumption cannot be generalized to all genes[48]. This means there may be tissue-specific signals we are not detecting. However, blood DNA methylation signal is an accessible diagnostic tool and DNA methylation sites from blood samples remain biologically relevant signals. Moreover, blood DNAm samples have been used to elucidate effects on other phenotypes, such as Alzheimer's and Type 2 Diabetes.[49-51] Another limitation is that both the UKBB and GoDMC sample predominantly European ancestries, which means there may be signals associated with non-European ancestry that are not being detected in this study. Thus, we plan to conduct a replicate analysis using the more genetically diverse biobank, All of Us.[52] Finally, we measure the warfarin average daily dose over the past 5 days, which may be a less robust metric than the patient's clotting time or time in the therapeutic range.[57]

While the longitudinal drug response model presented by Sadler *et al*. minimizes the risk of spurious signals unrelated to drug response,[10] it would still be useful to test whether any of the statin signals are replicated in a study of cholesterol levels alone. In addition, the signals identified in this study are directional from DNA methylation to the outcome of drug response. However, we know that some drugs and diseases induce DNA methylation changes. Therefore, it would be interesting to conduct an explicit bi-directional MR study to identify reverse-causal effects.[53] Moreover, we are learning that DNA methylation does not just regulate the nearest genes but has a more complex regulatory mechanism that may underlie these results.[3,24,31,32] Finally, it is difficult to compare effect sizes generated in this analysis with genetic effects identified through GWAS, because of the different assumptions and experimental set-ups. However, this work does demonstrate that epigenetic considerations are important for advancing our understanding of drug response and ADRs. In summary, we address the problem of insufficient and correlative studies linking DNA methylation and individual drug response with a statistical inference approach.

Acknowledgments

D.A.S. is supported by the NSF Graduate Research Fellowship 2019286895. S.A.A. is supported by the NIH T15 LM007033 from the NLM. M.C.S. was supported by the Swiss National Science Foundation (310030-189147). R.B.A. is supported by Chan-Zuckberg Biohub and NIH GM102365.

References

1. Kommu, S., Carter, C. & Whitfield, P. Adverse Drug Reactions. *StatPearls* (2024).
2. Chenchula, S., Atal, S. & Uppugunduri, C. R. S. A review of real-world evidence on preemptive pharmacogenomic testing for preventing adverse drug reactions: a reality for future health care. *Pharmacogenomics J* **24**, 1–6 (2024).
3. Smith, D. A., Sadler, M. C. & Altman, R. B. Promises and challenges in pharmacoepigenetics. *Camb Prism Precis Med* **1**, e18 (2023).
4. Yang, J. *et al*. Clopidogrel Resistance Is Associated With DNA Methylation of Genes From Whole Blood of Humans. *Front Genet* **11**, 583215 (2021).
5. Swen, J. J. et al. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomized crossover implementation study. The Lancet 401, 347-356 (2023).
6. Ochoa-Rosales, C. *et al*. Epigenetic Link Between Statin Therapy and Type 2 Diabetes. *Diabetes Care* **43**, 875–884 (2020).
7. Battram T., Yousefi P., Crawford G. *et al*. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res* vol. 7 (2022).
8. Sadler, M. C., Auwerx, C., Porcu, E. & Kutalik, Z. Quantifying Mediation between Omics Layers and Complex Traits. (2021) doi:10.1101/2021.09.29.462396.
9. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512–525 (2015).
10. Sadler, M. C. *et al*. Leveraging large-scale biobank EHRs to enhance pharmacogenetics of cardiometabolic disease medications. 2024.04.06.24305415 Preprint at https://doi.org/10.1101/2024.04.06.24305415 (2024).
11. Li, J., Wang, S., Barone, J. & Malone, B. Warfarin Pharmacogenomics. *P T* **34**, 422–427 (2009).
12. Scordo, M. G. *et al*. Influence of CYP2C9 and CYP2C19 genetic polymorphisms on warfarin maintenance dose and metabolic clearance. *Clin Pharmacol Ther* **72**, 702–710 (2002).
13. He, S. *et al*. Methylation of CYP1A1 and VKORC1 promoter associated with stable dosage of warfarin in Chinese patients. *PeerJ* **9**, 11549 (2021).
14. Zineh, I. Pharmacogenetics of Response to Statins. *Curr Atheroscler Rep* **9**, 187–194 (2007).
15. Florez, J. C. The pharmacogenetics of metformin. *Diabetologia* **60**, 1648–1655 (2017).
16. Cai, C., Wen, Z. & Li, L. The relationship between ApoE gene polymorphism and the efficacy of statins controlling hyperlipidemia. *Am J Transl Res* **13**, 6772–6777 (2021).
17. Zhou, K. *et al*. Variation in the glucose transporter gene SLC2A2 is associated with glycemic response to metformin. *Nat Genet* **48**, 1055–1059 (2016).
18. Marra, P. S. *et al.* Metformin use history and genome-wide DNA methylation profile: potential molecular mechanism for aging and longevity. *Aging (Albany NY)* **15**, 601–616 (2023).
19. Qin, X., Wang, Y., Pedersen, N. L., Tang, B. & Hägg, S. Dynamic patterns of blood lipids and DNA methylation in response to statin therapy. *Clin Epigenetics* **14**, 153 (2022).

20. Bycroft, C. *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
21. McInnes, G. & Altman, R. B. Drug Response Pharmacogenetics for 200,000 UK Biobank Participants. 2020.08.09.243311 Preprint at https://doi.org/10.1101/2020.08.09.243311 (2020).
22. Mbatchou, J. *et al*. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097–1103 (2021).
23. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet* **13**, 1007081 (2017).
24. Min, J. L. *et al*. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet* **53**, 1311–1321 (2021).
25. Sudmant, P. H. *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
26. Perktold, J. *et al.* statsmodels/statsmodels: Release 0.14.2. Zenodo https://doi.org/10.5281/zenodo.10984387 (2024).
27. Metagen R package (v4.9.6): Release 4.9.6. RDRR https://rdrr.io/cran/meta/src/R/metagen.R (2024).
28. R: The R Stats Package. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html (2024).
29. Sollis, E. *et al*. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, 977–985 (2023).
30. Bowden, J. *et al*. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International Journal of Epidemiology* **48**, 728–742 (2019).
31. Zaghlool, S. B. *et al*. Mendelian inheritance of trimodal CpG methylation sites suggests distal cis-acting genetic effects. *Clin Epigenetics* 8, 124 (2016).
32. Tong, Y. *et al*. MICMIC: identification of DNA methylation of distal regulatory regions with causal effects on tumorigenesis. *Genome Biol* 19, 73, (2018).
33. Pavlova, A., Zeitler, H., Scharrer, I., Brackmann, H.-H. & Oldenburg, J. HLA genotype in patients with acquired haemophilia A. *Haemophilia* **16**, 107–112 (2010).
34. Ivasková, E. *et al*. HLA and venous thrombosis: a prospective study. *Cor Vasa* **33**, 424–427 (1991).
35. Laghmouchi, A., Graça, N. A. G. & Voorberg, J. Emerging Concepts in Immune Thrombotic Thrombocytopenic Purpura. *Front Immunol* **12**, 757192 (2021).
36. Wong, K., Lai, W. K. & Jackson, D. E. HLA Class II regulation of immune response in sickle cell disease patients: Susceptibility to red blood cell alloimmunization (systematic review and meta-analysis. *Vox Sang* **117**, 1251–1261 (2022).
37. Whirl-Carrillo, M. *et al*. An Evidence-Based Framework for Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clin Pharmacol Ther* **110**, 563–572 (2021).
38. Zorzo, R. A. *et al*. LDLR gene's promoter region hypermethylation in patients with familial hypercholesterolemia. *Sci Rep* **13**, 9241 (2023).

39. Behbodikhah, J. *et al*. Apolipoprotein B and Cardiovascular Disease: Biomarker and Potential Therapeutic Target. *Metabolites* **11**, 690 (2021).
40. Richardson, T. G. *et al*. Characterising metabolomic signatures of lipid-modifying therapies through drug target mendelian randomisation. *PLoS Biol* **20**, 3001547 (2022).
41. Jia, X., Zhao, C. & Zhao, W. Emerging Roles of MHC Class I Region-Encoded E3 Ubiquitin Ligases in Innate Immunity. *Front. Immunol* **12**, (2021).
42. Sakaue, S. *et al*. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415–1424 (2021).
43. Li, S. *et al*. Interaction of SH3P13 and DYDC1 protein: a germ cell component that regulates acrosome biogenesis during spermiogenesis. *Eur J Cell Biol* **88**, 509–520 (2009).
44. Ventham, N. T. *et al*. Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat Commun* **7**, 13507 (2016).
45. Boyalla, V., Gallego-Colon, E. & Spartalis, M. Immunity and inflammation in cardiovascular disorders. *BMC Cardiovasc Disord* **23**, 148 (2023).
46. Dal Lin, C., Tona, F. & Osto, E. The crosstalk between the cardiovascular and the immune system. *Vasc Biol* **1**, 83–88 (2019).
47. Wielscher, M. *et al*. DNA methylation signature of chronic low-grade inflammation and its role in cardio-respiratory diseases. *Nat Commun* **13**, 2408 (2022).
48. Olsson Lindvall, M. *et al*. Comparison of DNA Methylation Profiles of Hemostatic Genes between Liver Tissue and Peripheral Blood within Individuals. *Thromb Haemost* **121**, 573–583 (2021).
49. Konki, M. *et al*. Peripheral blood DNA methylation differences in twin pairs discordant for Alzheimer's disease. *Clinical Epigenetics* **11**, 130 (2019).
50. Sun, Y. *et al*. Identification of candidate DNA methylation biomarkers related to Alzheimer's disease risk by integrating genome and blood methylome data. *Transl Psychiatry* **13**, 1–10 (2023).
51. Juvinao-Quintero, D. L., Sharp, G. C., Sanderson, E. C. M., Relton, C. L. & Elliott, H. R. Investigating causality in the association between DNA methylation and type 2 diabetes using bidirectional two-sample Mendelian randomisation. *Diabetologia* **66**, 1247–1259 (2023).
52. Mayer, C. S. & Huser, V. Learning important common data elements from shared study data: The All of Us program analysis. *PLoS One* **18**, 0283601 (2023).
53. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, 89–98 (2014).
54. Stadler, M. B. et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480, 490–495 (2011).
55. Glavinovic, T. et al. Physiological Bases for the Superiority of Apolipoprotein B Over Low-Density Lipoprotein Cholesterol and Non–High-Density Lipoprotein Cholesterol as a Marker of Cardiovascular Risk. J Am Heart Assoc 11, e025858 (2022).
56. Vega, G. L. & Grundy, S. M. Hypercholesterolemia with cholesterol-enriched LDL and normal levels of LDL-apolipoprotein B. Effects of the step I diet and bile acid sequestrants on the cholesterol content of LDL. Arterioscler Thromb Vasc Biol 16, 517–522 (1996).

57. Kuruvilla, M. & Gurk-Turner, C. A review of warfarin dosing and monitoring. Proc (Bayl Univ Med Cent) 14, 305–306 (2001).