

A comparison of the genes and genesets identified by GWAS and EWAS of fifteen complex traits

Received: 23 February 2022

Accepted: 16 November 2022

Published online: 19 December 2022

 Check for updates

Thomas Battram ^{1,2} ✉, Tom R. Gaunt ^{1,2}, Caroline L. Relton ^{1,2},
Nicholas J. Timpson ^{1,2} & Gibran Hemani ^{1,2}

Identifying genomic regions pertinent to complex traits is a common goal of genome-wide and epigenome-wide association studies (GWAS and EWAS). GWAS identify causal genetic variants, directly or via linkage disequilibrium, and EWAS identify variation in DNA methylation associated with a trait. While GWAS in principle will only detect variants due to causal genes, EWAS can also identify genes via confounding, or reverse causation. We systematically compare GWAS ($N > 50,000$) and EWAS ($N > 4500$) results of 15 complex traits. We evaluate if the genes or gene ontology terms flagged by GWAS and EWAS overlap, and find substantial overlap for diastolic blood pressure, (gene overlap $P = 5.2 \times 10^{-6}$; term overlap $P = 0.001$). We superimpose our empirical findings against simulated models of varying genetic and epigenetic architectures and observe that in most cases GWAS and EWAS are likely capturing distinct genesets. Our results indicate that GWAS and EWAS are capturing different aspects of the biology of complex traits.

Determining molecular associations with complex traits can yield novel therapeutic interventions. Large efforts have been made to conduct hypothesis-free searches to identify associations between disease variation and variation in molecular features. Two popular study designs include genome-wide association studies (GWAS) and epigenome-wide association studies (EWAS). GWAS and EWAS identify mechanistically distinct aspects of inter-individual genomic variation which may lead them to prioritise different aspects of disease biology. In this paper, we investigate the relationship between the results of these two approaches.

EWAS assess the association between one facet of epigenetics, DNA methylation (DNAm), and complex traits^{1–3}. Often in EWAS, the potential biological implications of differentially methylated positions or regions (DMPs or DMRs) will be investigated further through genomic annotations^{4–7}. Previous studies have demonstrated a relationship between DNAm levels and proximal genes^{8,9}. This observation has led to it being common place to map sites identified in EWAS to nearby genes and these genes and their functions are often probed to ascertain their relevance to the trait of interest^{4–7}. Further, genes can

be grouped with others into “genesets” that have similar functionality or lie within the same biological pathway. Examining over-represented genesets may provide an insight into the molecular biology of a trait. For example, Reese et al. (2019) performed an EWAS of asthma and discovered the DMPs and DMRs identified mapped to genes within relevant immunological pathways more than expected by chance⁵. These genesets included endothelial nitric oxide synthase (eNOS) signalling, the inflammasome, and nuclear factor κ B (NF- κ B) signalling⁵. Other assessments may be made to attempt to infer biological understanding, including enrichment of other epigenetic marks at the regions identified¹⁰ and follow-up experimental studies¹¹. However, using open access databases to investigate tagged genes and genesets is a simple and potentially effective approach to further biological understanding.

GWAS often use similar approaches to EWAS to help in the discovery of genomic regions related to complex traits¹². Examples of genetic-epigenetic equivalence are known, that is the identical, clinically measurable, effects of a genetic lesion and an epigenetic change. For example, Angelman syndrome and Prader Willi syndrome can both

¹MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. ²Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK.

✉ e-mail: thomas.battram@bristol.ac.uk

be caused by a deletion mutation or by imprinting^{13,14}. However, the properties of genetic variants and DNAm differ in terms of the causal mechanisms that can give rise to trait associations, making potential inferences from EWAS and GWAS diverge. Importantly, DNAm is responsive to environmental stimuli, thus making associations identified in EWAS potentially attributable to forward causation (where the DNAm level is causal for the trait), reverse causation (where the trait influences the DNAm level) and to confounding (where one factor influences both DNAm level and trait)^{2,3,15}. Recent work has suggested that there is a general trend for DNAm associations with complex traits being more likely due to confounding or reverse causation than DNAm itself being causal^{16,17}. A simple theoretical model suggests that statistical power in EWAS to detect associations due to reverse causation far exceeds those due to forward causation or confounding (Supplementary Information - Supplementary Note). It should be noted that GWAS may be susceptible to confounding due to various population phenomena such as population stratification or dynastic effects¹⁸, but statistical adjustments are routinely applied to address this, and confounding is unlikely to be affecting a majority of GWAS signals, especially for clinical traits^{19,20}.

A direct comparison between GWAS and EWAS results could provide insight into what biological information EWAS are capturing. If EWAS are highlighting a similar set of genes and genesets to GWAS, it suggests changes in DNAm are either themselves involved in trait aetiology or tagging molecular aetiologically relevant genomic features. Similarly, DMPs identified as the result of confounding and reverse causation may also highlight similar genes and genesets as GWAS if these confounding and reverse causal pathways are similar to the causal pathways for a trait. Regardless, in that scenario EWAS is still identifying facets of trait aetiology. Under the circumstance that biological insights from GWAS and EWAS overlap, the efficiency of EWAS would far outstrip that of GWAS, for example an EWAS of BMI using around 10,000 samples identified 187 independent genomic loci¹⁷ while it required a GWAS of 330,000 samples to identify 97 independent genomic loci²¹. Yet, via Mendelian randomisation analyses, Wahl et al. (2017) provided evidence that BMI likely caused a change in DNAm at the majority of these 187 loci.

In the event that GWAS and EWAS are not highlighting a similar set of genes and genesets, it is plausible that EWAS may still be identifying facets of trait aetiology. If DNAm mediates non-genetic effects or if sites are mapped to genes or genesets incorrectly then overlap between highlighted genes and genesets will not be guaranteed even if

DNAm changes identified are aetiologically relevant. When DNAm mediates the effect of genetic variants distal to their genomic position on complex traits the genes identified by GWAS and EWAS will also differ, but the genesets would likely overlap. A lack of overlap could also reflect that associations in EWAS may be driven by confounding and reverse causation. Despite the caveats mentioned, it is plausible that in the absence of confounding and reverse causation genes and genesets identified in EWAS would overlap more than expected by chance with those identified by GWAS. The extent to which this expectation holds is explored in more detail in the “Discussion”.

In this paper, we determine the overlap between genes and genesets identified by GWAS and EWAS of 15 complex traits and we use these results to infer genetic and epigenetic architectures that are likely to be consistent with our results. We do not attempt to infer the causal structure of DNAm-trait associations directly as has been attempted through methods such as Mendelian randomisation, but is often difficult due to being unable to prove the vertical pleiotropy assumption^{16,22}. Rather, we adopt a different approach by simulating a range of causal systems to explore potential interpretations of the patterns of overlap between genomic features reported by GWAS and EWAS. For convenience, we denote the terms “causal gene” as a gene that influences the trait of interest and an “associated gene” as a gene that correlates with the trait of interest, without necessarily being causal (Fig. 1).

Results

Study data

We selected traits for which EWAS had been conducted with over 4500 samples, had more than 10 associations at $P < 1 \times 10^{-7}$ and for which corresponding well-powered GWAS summary data were available. Suitable traits were identified using The EWAS Catalogue²³ on 2021-12-20. Traits and study data information is in Table 1.

Genomic position overlap

We divided the genome into 5591 non-overlapping 500 kb regions that each contained at least one probe from the Illumina 450k array, and mapped GWAS and EWAS signals for each trait to these regions (see “Methods” for more details). For each trait, the number of regions that were identified by one study type and not the other was higher than the number of overlapping regions (Fig. 2). Further, the magnitude of the greatest GWAS effect estimate in each region had little ability to predict whether or not a DNAm site was likely to be identified in the same region (AUC range = 0.43–0.61, Supplementary Fig. 1).

Assessing power to detect shared annotations between GWAS and EWAS

Genomic function and trait biology are not divided into discrete 500 kb genomic chunks, thus GWAS and EWAS could still be identifying similar facets of trait biology without identifying the same genomic regions.

We sought to assess whether the genes and genesets identified overlapped more than expected by chance, thus genomic positions were mapped to genes and genes to genesets (details in “Methods”). Overlap between genes identified was assessed using Fisher’s exact test. Two methods for testing overlap between genesets were considered, one simply mapped genes to genesets and used Fisher’s exact test in the same way as assessing the overlap between identified genes. The other generated ‘enrichment scores’ for each geneset and assessed the correlation between the geneset enrichment scores across studies.

In the case that genetic influences on disease are mediated via cis-acting DNAm levels, or other proximal regulatory factors that in turn correlate with proximal DNAm levels, if all DMPs associated with a trait were on the pathway from SNP to disease, then GWAS and EWAS would be identifying genes from the exact same geneset (i.e., genes that

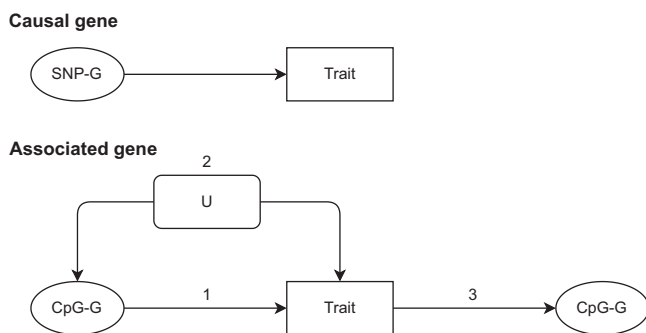


Fig. 1 | A diagram of causal and associated genes. A causal gene is one where the product of that gene affects the trait of interest and in this study, we are assuming that SNPs identified in relation to a trait will affect these genes or tag SNPs that do (SNP-G). An associated gene is one where the product of that gene correlates with the trait of interest, but may not affect it. In this study, we are assuming that CpG sites identified in relation to a trait will map to these genes (CpG-G). The diagram shows how a gene product may be correlated with a trait: 1. by affecting the trait, 2. by sharing a common cause with the trait (confounding), 3. by being affected by the trait (reverse causation). A geneset may be composed of causal and associated gene products. U = confounder.

Table 1 | Study data

Trait	EWAS author	EWAS PMID	EWAS <i>n</i>	GWAS author	GWAS PMID	GWAS <i>n</i>
Body mass index	Wahl S	28002404 ¹⁷	10,238	Yengo L	30124842 ⁶⁵	681,275
Current versus never smoking	Joehanes R	27651444 ⁷⁵	9389	Liu M	30643251 ⁷⁶	632,802
Former versus never smoking	Joehanes R	27651444 ⁷⁵	13,474	Elsworth B	NA	424,960
Alcohol consumption per day	Liu C	27843151 ⁷⁷	9643	Liu M	30643251 ⁷⁶	335,394
C-reactive protein	Ligthart S	27955697 ⁷⁸	8863	Ligthart S	30388399 ⁷⁹	204,402
Educational attainment	Karlsson Linner R	29086770 ⁸⁰	10,767	Lee JJ	30038396 ⁸¹	766,345
Fasting glucose	Liu J	31197173 ⁷⁶	4808	Manning AK	22581228 ⁸²	58,074
Fasting insulin	Liu J	31197173 ⁷⁶	4808	Manning AK	22581228 ⁸²	51,750
Systolic blood pressure	Richard MA	29198723 ⁵⁰	17,010	Evangelou E	30224653 ⁸³	757,601
Diastolic blood pressure	Richard MA	29198723 ⁵⁰	17,010	Evangelou E	30224653 ⁸³	757,601
Birthweight	Kupers L	31015461 ⁸⁴	8825	Horikoshi M	27680694 ⁸⁵	143,677
Cognitive abilities: digit test	Marioni R	29311653 ⁸⁶	4794	Lee JJ	30038396 ⁸¹	257,841
FEV1	Imboden M	31073081 ⁸⁷	5370	Elsworth B	NA	421,986
eGFR	Schlosser P	34887417 ⁸⁸	33,605	Stanzick KJ	34272381 ⁸⁹	961,734
Urate	Tin A	34887389 ⁵²	17,996	Tin A	31578528 ⁹⁰	278,592

PMID PubMed ID, author = first author on the publication. References can be found after each PMID.

Where GWAS PMID = NA, the GWAS were conducted as part of a UK Biobank GWAS pipeline within the University of Bristol's Integrative Epidemiology Unit and can be found on the OpenGWAS Project website (see "Methods" for more).

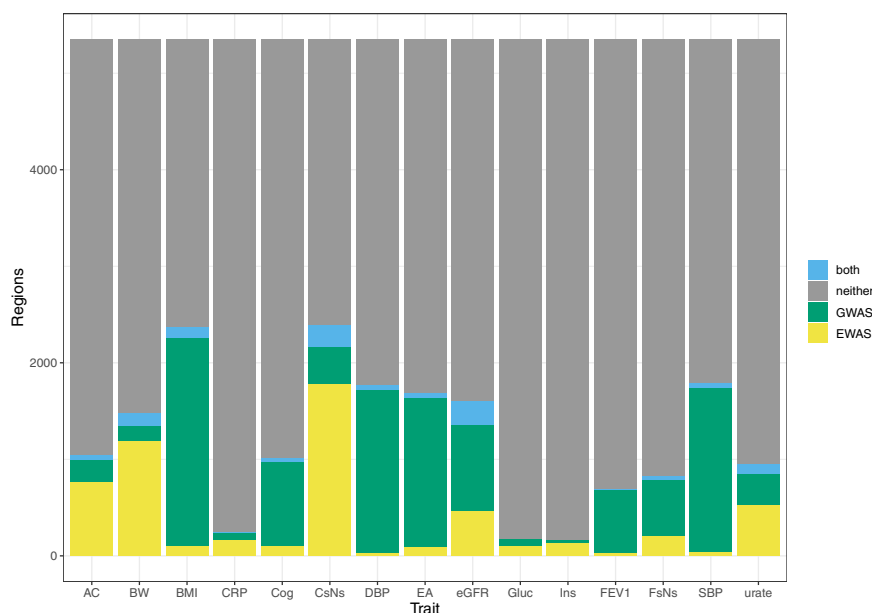


Fig. 2 | Overlap between genomic positions identified by corresponding GWAS and EWAS. The genome was divided into 500kb regions. Those where no probes on the HM450 array measured DNAm were excluded from the analysis. This left 5591 regions. Regions were counted as being identified by a GWAS if one or more SNPs in that region associated with the trait and as being identified by an EWAS if one or more CpGs in that region associated with the trait. Neither = no GWAS or EWAS sites identified in the region, GWAS = GWAS sites only were identified,

EWAS = EWAS sites only were identified, Both = Both GWAS and EWAS sites were identified, AC alcohol consumption per day, BW birthweight, BMI body mass index, Cog cognitive ability (digit test), CRP c-reactive protein, CsNs current smokers vs never smokers, DBP diastolic blood pressure, EA educational attainment, Gluc fasting glucose, Ins fasting insulin, FEV1 forced expiratory volume in one second, FsNs former smokers vs never smokers, SBP systolic blood pressure.

caused changes in the trait). We note that there are ways in which EWAS might be identifying aetiologically relevant signal without being at the same loci as GWAS signal, which is discussed in detail in the Introduction and Discussion. The more DMPs that are identified because of confounding effects or reverse causation, the smaller the chance of overlap, assuming that causal and responsive genesets are independent (more on this in the "Discussion"). In a scenario where no DMPs are causing phenotypic changes then any overlap in genes and

genesets found would be entirely attributable to chance. We ran simulations to assess which scenarios the enrichment and annotation methods had power to detect whether there was more overlap than expected by chance. Power was also assessed across different annotation methods. Box 1 shows the steps involved in this simulation.

Under each of a range of genetic and epigenetic architectures and study sizes, the ability to infer whether EWAS were identifying, at least in part, the same set of genes as GWAS ('causal genes') compared to a

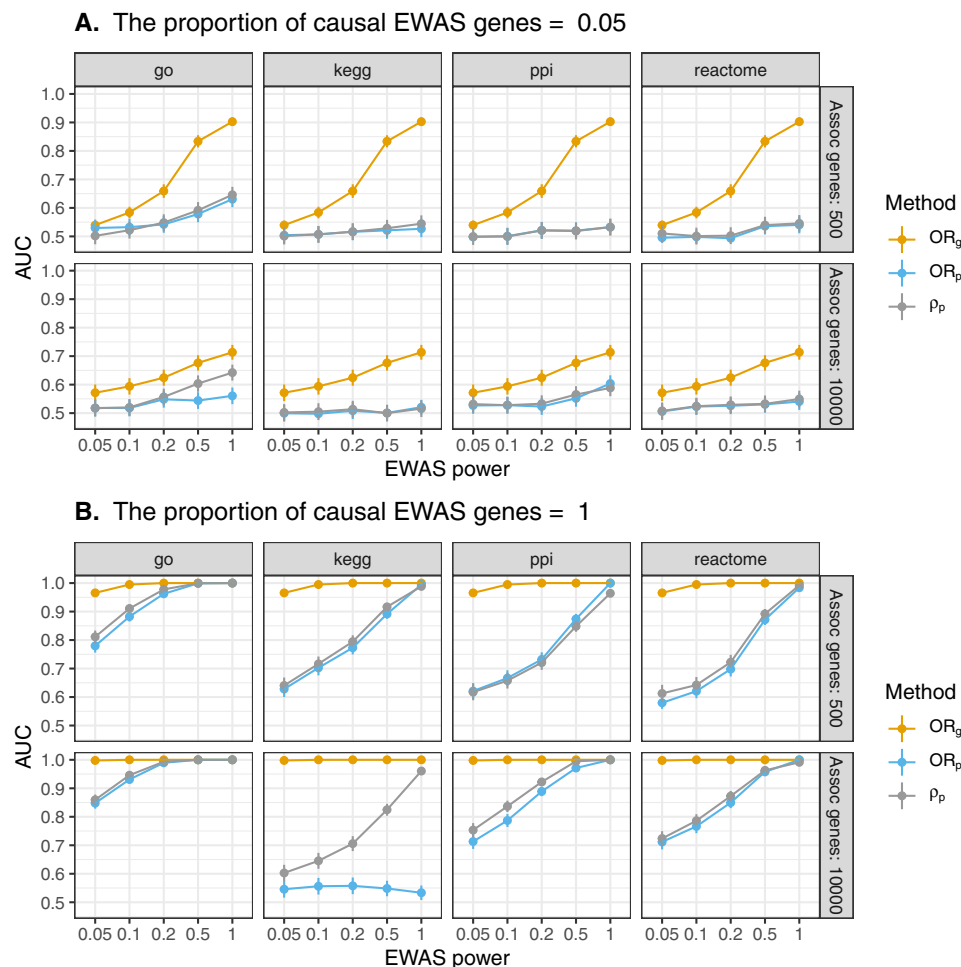


Fig. 3 | Power to detect overlap between genes and genesets identified by corresponding GWAS and EWAS. Simulations were set up as illustrated in Box 1 and simulations iterated over each set of parameters 1000 times. EWAS power is equivalent to the proportion of associated genes (Assoc genes) EWAS is detecting. In the scenario where Assoc genes = 500, EWAS power = 1, and the proportion of causal EWAS genes = 0.05, the EWAS is detecting 500 genes, 25 of which are causal. Panel **A** show results when the proportion of causal EWAS genes = 0.05 and panel **B** show results when the proportion of causal EWAS genes = 1. The area under receiver operator curves (AUC) was used to estimate the ability to distinguish

between results generated when GWAS and EWAS were sampling, in part, from the same set of causal genes and results generated when EWAS was sampling random genes from the genome. Error bars represent the 95% confidence intervals of the AUC estimates. The header of each set indicates the proportion of genes identified by the simulated EWAS that were set to be causal. OR_g = assessing overlap of genes, OR_p = assessing overlap of genesets, ρ_p = assessing correlation between geneset enrichment scores. GO gene ontology, PPI protein–protein interaction database from EpiGraphDB. This is a summary of the results, full results can be found in Supplementary Fig. 3.

random set of genes (‘associated genes’) was tested. Performance improved as the study sample sizes and the proportion of DMPs that were causal increased (Fig. 3 and Supplementary Fig. 3). Performance of the overlap tests tended to increase as the number of identified genes increased, but this parameter was largely inconsequential when the proportion of DMPs that were causal was low (Supplementary Fig. 3). Performance was similar across annotation methods and between methods attempting to assess geneset overlap, with assessment of gene overlap performing better (Fig. 3). Overall there was more power to detect overlap in genes than overlap in genesets. Between the geneset methods, there was more power to detect a correlation between enrichment scores than direct overlap in genesets. Therefore, gene overlap and correlation between geneset enrichment scores were taken forward for the empirical analyses. Assigning genes to GO terms^{24,25} to use as genesets had more power, than assigning genes to genesets derived from the KEGG^{26–28}, Reactome²⁹ or protein–protein interaction database from EpiGraphDB³⁰, when assessing overlap between these genesets (Fig. 3). Therefore, GO terms were taken forward and used as the geneset annotations in the empirical analyses.

Gene and geneset overlap between GWAS and EWAS

For the 15 traits used, the number of genes identified by GWAS and EWAS that overlapped was low and for three traits no genes identified by the studies overlapped (Table 2). The number of genesets that overlapped was higher, peaking at 1264 for GWAS and EWAS of eGFR (Table 3).

The number of overlapping genes identified was no more than expected by chance for 11 of 15 traits, but for systolic blood pressure, diastolic blood pressure, eGFR, and urate the overlap between observed genes was greater than expected (6, 9, 24, and 12 more genes overlapped than expected by chance respectively) (Table 2). There was also evidence that correlation between enrichment scores of the GO terms identified by GWAS and EWAS of diastolic blood pressure was greater than expected by chance ($P = 0.0013$), but there was little evidence for this across the other traits, including systolic blood pressure, egr, and urate (Table 3).

There were 56 GO term genesets that were commonly enriched (FDR < 0.1) for both the GWAS and EWAS traits. All of these genesets were relatively broad (contained over 100 genes). There were 261 genesets of size < 100 genes that did not overlap between studies of

Table 2 | Overlap of genes identified by EWAS and GWAS

Trait	<i>n</i> EWAS genes	<i>n</i> GWAS genes	Gene overlap	obs OR	exp overlap	exp OR	<i>p</i> diff
Body mass index	232	3228	23	2.0	39	3.04	5.2e-02
Current versus never smoking	1933	312	27	2.9	39	3.42	3.2e-01
Former versus never smoking	282	323	9	6.4	7	3.43	2.2e-02
Alcohol consumption per day	361	190	3	2.6	3	2.44	9.1e-01
C-reactive protein	189	302	3	3.2	1	0.91	1.2e-01
Educational attainment	25	1594	1	1.5	3	3.00	4.3e-01
Fasting glucose	15	50	0	0.0	0	0.00	1.0e+00
Fasting insulin	36	5	0	0.0	0	0.00	1.0e+00
Systolic blood pressure	99	3210	18	4.0	12	2.13	9.4e-03
Diastolic blood pressure	56	3586	15	5.8	6	1.87	5.2e-06
Birthweight	894	162	6	2.6	7	2.64	9.1e-01
Cognitive abilities: digit test	26	875	0	0.0	1	2.44	4.2e-01
FEV1	51	1664	4	3.0	3	1.92	3.9e-01
eGFR	256	2611	44	4.7	20	1.81	2.1e-10
Urate	251	666	16	6.3	4	1.28	3.5e-11

exp = expected, obs = observed, fev1 = forced expiratory volume in 1 s.

odds ratios (ORs) can be interpreted as the odds of a gene being identified by EWAS and a GWAS over the odds of a gene being identified by an EWAS but not by a GWAS.

exp OR = the mean OR after repeating the analysis 1000 times, randomly sampling EWAS genes equal to the number identified in the empirical analysis.

p diff = *P* value from a two-sided Z-test assessing the difference between the observed and expected OR.

Table 3 | Correlation of geneset enrichment scores between EWAS and GWAS

trait	<i>n</i> EWAS genes	<i>n</i> GWAS genes	Geneset overlap	obs cor	exp cor	<i>p</i> diff
Body mass index	232	3,228	1,243	0.187	0.20	0.2287
Current versus never smoking	1933	312	1,053	0.200	0.22	0.0343
Former versus never smoking	282	323	661	0.298	0.30	0.9936
Alcohol consumption per day	361	190	562	0.259	0.27	0.5753
C-reactive protein	189	302	600	0.265	0.26	0.7983
Educational attainment	25	1,594	215	0.105	0.13	0.1729
Fasting glucose	15	50	50	0.153	0.16	0.7723
Fasting insulin	36	5	19	0.093	0.12	0.4400
Systolic blood pressure	99	3,210	631	0.145	0.14	0.7679
Diastolic blood pressure	56	3586	594	0.160	0.11	0.0013
Birthweight	894	162	683	0.224	0.24	0.2476
Cognitive abilities: digit test	26	875	198	0.164	0.14	0.2064
FEV1	51	1,664	403	0.158	0.14	0.1948
eGFR	256	2611	1,264	0.233	0.20	0.0332
Urate	251	666	877	0.306	0.27	0.0414

For each geneset, odds of study genes being in the geneset divided by the odds of the study genes not being in the geneset were assessed and correlation between these odds ratios are given here.

exp cor = the mean correlation between odds ratios after repeating the analysis 1000 times, randomly sampling EWAS genes equal to the number identified in the empirical analysis.

geneset overlap indicates the number of gene ontology terms that map to both genes identified by the EWAS and GWAS.

p diff = *P* value from a two-sided Z-test assessing the difference between the observed and expected correlations.

corresponding traits, for example, the genes identified by the GWAS of alcohol consumption were enriched for the “ethanol catabolism” pathway (geneset size = 12 genes), however none of the genes identified by the EWAS were present in this pathway.

Understanding architecture from geneset overlap

Given observations of numbers of genesets discovered and geneset overlap for a particular trait, we next asked if we can impose bounds on the likely genetic and epigenetic architectures of that trait. To do this we assumed discovered genes and genesets represented some fraction of total genes and genesets, and then re-sampled GWAS and EWAS results to ascertain architectures that generated overlap scores matching the empirical results.

For the simulations, three sets of genes were linked to each trait: genes identified by the GWAS (known GWAS genes), genes identified by the EWAS (known EWAS genes), and a random set of genes sampled from the total set of Ensembl gene IDs (excluding the genes identified by the GWAS and EWAS).

Having generated the GWAS and EWAS genes, enrichment of GO terms was performed and the correlation between enrichment scores across all the terms was estimated. A schematic of the methods for these simulations can be found in Supplementary Fig. 4 and it is described in full in the “Methods”.

From the simulations, we determined, that for many traits the overlap between causal and associated genes is unlikely to be high if the total number of genes still to discover for these traits is low. However,

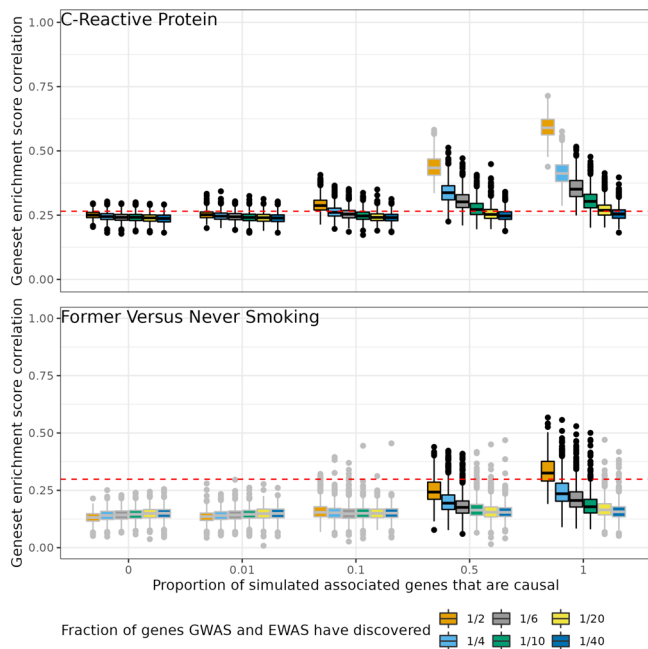


Fig. 4 | Simulations to understand the likely number of genes still to identify in GWAS and EWAS of C-reactive protein and smoking (former vs. never smokers) under different trait architectures. Simulations were set up as illustrated in Supplementary Fig. 4. Correlation of geneset enrichment scores from empirical data (Table 3), is shown as a red dashed line. Box plots show the range of enrichment score correlations from 1000 simulations using the parameters indicated. The number of causal and associated genes, as well as the number of associated genes that were causal were varied. Already discovered EWAS genes were added to the pool of associated genes and already discovered GWAS genes were added to the pool of causal genes. The proportion of simulated associated genes that were causal is shown on the X-axis. The number of causal genes and associated genes were equal in each simulation. Scenarios which lie close to the empirical result (red dashed line) are more likely to reflect the true underlying number of genes related to a trait and the true overlap between the causal and associated genes. Where there is evidence that on average the geneset enrichment scores from a simulation scenario are different to the empirical enrichment score ($FDR < 0.05$, z-test for difference), the box outline is grey, otherwise it is black. The centre of the box plots are the median, the bounds of the box represent the interquartile range (IQR), the upper whisker represents either the minimum of $(1.5 \times \text{IQR}) + \text{the } 75\% \text{ percentile}$ and the maximum value, the lower whisker represents the maximum of $25\% \text{ percentile} - (1.5 \times \text{IQR})$ and the maximum value. Values that fall outside the whiskers are marked as points.

for some traits such as systolic blood pressure, diastolic blood pressure, and former vs. never smoking the simulations suggest the observed degree of overlap corresponds to a high overall overlap between causal and associated genes. It should be noted though, that the correlation of enrichment scores across different scenarios varied little for each trait, making inference difficult. The results from the analysis of former vs. never smoking and C-reactive protein (representing simulations for the other traits) are shown in Fig. 4. Supplementary Figure 5 shows the results for the other traits. Of 405 scenarios, there was some evidence against 124 reflecting reality ($FDR < 0.05$). Across the traits, the scenarios that were least likely tended to be when the number of genes yet to discover was low, and the overlap between causal and associated genes was high, except for former vs. never smoking, highlighting architecture differences between traits.

Overlap of non-corresponding GWAS and EWAS

We next hypothesised that EWAS of one trait might actually relate more closely to GWAS of another trait. For example, if DNAm changes related to BMI were mediating the effect of BMI on changes in

metabolites, then one could expect to see greater correspondence in identified genes and genesets with GWAS of those metabolites.

To test if this was the case for any of the EWAS in this study we extracted 1886 GWAS from the OpenGWAS Project^{31,32}. The criteria for inclusion can be found in the Methods. The GWAS extracted consisted of a variety of traits, including diseases like breast cancer, anthropometric measures, measures of cognitive performance, metabolite measures. All traits can be found in Supplementary Data 1. For each GWAS and the 15 EWAS, enrichment scores were calculated using geneset enrichment analysis, and correlation between the enrichment scores was calculated, as in previous analyses.

Across all pairwise comparisons, correlations between enrichment scores ranged from -0.014 to 1 and had a mean of 0.12 (Supplementary Fig. 6). The mean correlation between GWAS traits and EWAS traits was 0.26, which was higher than the correlations between just GWAS traits (0.12). Amongst just GWAS results, there was evidence that 14652 pairwise enrichment score correlations were greater than the the mean ($FDR < 0.05$). However, there was little evidence that any pairwise correlations between enrichment scores derived from GWAS and EWAS were greater than the mean correlation ($FDR > 0.05$).

This suggests that the signal from EWAS is not capturing aspects of any specific factor that impacts the aetiology of any of the 15 traits of interest.

Discussion

Several EWAS papers have compared their findings with those of the corresponding trait GWAS^{6,33–36}, but it is unknown if any overlap that might occur should be attributed to shared underlying architectures or if it occurs by chance. In this study, the genes identified by 11 of 15 large EWAS ($N > 4500$) were not identified in their corresponding GWAS any more than expected by chance, and only one EWAS identified genesets overlapping with the corresponding GWAS more than expected by chance. Simulations suggested the other EWAS could still be identifying aspects of trait aetiology but given EWAS likely have greater power to detect reverse causal associations (Supplementary Information) and prior evidence for EWAS findings not being causal^{16,17}, it seems likely most DMPs identified are due reverse causation. Further simulations suggested that the overlap between genes that impact phenotypic variation and those that might be identified through confounded analyses or because of reverse causation is likely to be low amongst EWAS of most traits. However, if the number of genes still to identify in GWAS and EWAS is high (e.g., if we have discovered less than one quarter so far), it is possible that the number of associated genes discovered by EWAS that are causal could be high (over 50%).

Overlap expected

GWAS identifies the effects of genetic variation on complex traits. These effects are less likely to be confounded than associations estimated between observational phenotypes^{37,38}. Thus, one would expect overlap between genes and genesets identified by GWAS and EWAS of the same trait if the DMPs identified are also of aetiological relevance. Assuming mapping of DMPs and SNPs to genes is correct, the genes identified by EWAS may cause variation in complex traits without overlapping with genes identified in GWAS. Under the scenario where the effect of a SNP on a complex trait is mediated by a distal DNAm site (or the gene that site is tagging), GWAS and EWAS may identify genes that do not overlap, but that are along the same causal pathway from genome to complex trait. One plausible mechanism for which transmethylation quantitative trait loci (trans-meQTLs; that is SNPs that effect DNAm at distal sites) may act is via transcription factors. A trans-meQTL may influence the transcription of a nearby transcription factor then the transcription factor could cause a change in DNAm at distal sites. One study has provided evidence this may occur frequently³⁹. In this scenario the genes proximal to the identified SNP and DMP would lie on the same causal pathway and so would likely be part of the same

genesets. This is likely not the only plausible mechanism of trans-meQTL function though. DNAm may also mediate non-genetic effects, allowing for causal DMPs to be identified at genes not near pertinent genetic variation. However, there is strong evidence that the majority of DNAm sites have a heritable component^{16,40,41}. As such any effect of DNAm on a trait could be influenced by genetic variation. If DNAm is influenced by proximal genetic variants then the discovery of the same gene(s) will be a function of GWAS, EWAS power and the heritability of the DNAm sites. If only distal genetic variants influence the DNAm site, then the overlap of genesets is a function of regulatory mechanisms and power. These two issues, and likely others, may introduce some noise into the results. However, there is overwhelming evidence that confounding and reverse causation are pervasive across observational epidemiology^{42–45} as well as within EWAS^{3,46}. This suggests that the evidence that GWAS and EWAS are not identifying any more overlapping genes and genesets than expected by chance, is likely due to identified DMPs mostly being the result of reverse causation or confounding, and that these different mechanisms have relatively distinct biological functions.

As the simulations showed, even if an EWAS identifies DMPs that cause a change in the trait, if the majority of DMPs identified are due to confounding or reverse causation then the overlap will be indistinguishable to the overlap expected by chance. Thus, our empirical results do not preclude the possibility that some DMPs identified by the EWAS are involved in trait aetiology.

For body mass index, work has already suggested the trait causes changes in DNAm rather than vice versa¹⁷, supporting our findings here. Further, one study suggested DNAm changes capture different components of body mass index variance than genetic variation⁴⁷ and another estimated the percentage of trait variance captured by DNAm was 76% when accounting for genotype⁴⁸.

Simulations involving empirical data from former vs. never smoking GWAS and EWAS suggested that overlap between causal and associated genes was high. This is surprising for two reasons. Firstly, it differs from the current vs. never smoking results, suggesting distinct genetic or epigenetic architectures of those traits. Secondly, there is evidence that for DNAm changes identified in relation to smoking, smoking is likely causing DNAm variation and not vice versa⁴⁹. Although the statistical tests suggested that it was unlikely that the proportion of causal and associated genes is 0.1 or lower, it should be noted that the absolute difference between simulated results and empirical results were subtle. Another consideration is that 'former smoking status' is a progression measure, which makes the former vs never smoking variable susceptible to collider bias. Throughout this work we have not investigated how collider bias might impact gene or geneset enrichment, and it is possible that this has an impact on the results.

It's also potentially unexpected that genes identified by EWAS diastolic and systolic blood pressure overlapped more with genes identified by their respective GWAS than expected by chance. In the study that conducted the EWAS of both blood pressure phenotypes, the authors conducted bi-directional Mendelian randomisation analyses to improve causal inference of DNAm changes identified in the EWAS⁵⁰. They presented evidence that variation in blood pressure causes changes in DNAm at 4 sites, and evidence for the reverse at one site. However, not all CpG sites could be instrumented and not all sites identified in the EWAS were taken forward for the Mendelian randomisation analyses, with more analyses being run to ascertain the effect of blood pressure on DNAm than the reverse. Further, for the sites taken forward, none had a large number of genetic instruments, making evaluation of pleiotropy difficult. Therefore, it is plausible that some of the sites identified in their EWAS are tagging causal genes and this could not be tested by the authors. In addition to this, a GWAS of blood pressure found that identified SNPs were enriched for association with DNAm at CpG sites within IMb⁵¹. Similar, to our study, this

work also suggests that DNAm variation is tagging causal blood pressure loci, without formally testing whether DNAm changes affect the trait directly.

The latest urate EWAS is potentially in contrast with other EWAS, whereby the authors find evidence that five CpGs are associated with urate at the locus with the strongest GWAS signal, *SLC2A9*, and that two of these CpGs mediate some of the effect genetic variants have on urate levels⁵². This makes it unsurprising that we found strong evidence the overlap between genes in urate EWAS and GWAS is greater than expected by chance.

If EWAS is discovering some DMPs that influence genes that cause changes in the trait of interest, study power is the limiting factor for detecting overlap. For traits with a weak polygenic architecture (few genes explain most of the heritability), such as gene expression^{53,54}, discovering almost total overlap might be possible even with modest sample sizes.

Little overlap with any GWAS

Little correlation was found between geneset enrichment scores for GWAS and EWAS of different traits. In a scenario where DNAm was capturing a specific facet of a trait one might expect a correlation between EWAS of the original trait and GWAS of that facet. For example, if changes in DNAm associated with smoking were mostly responsible for the effect of smoking on lung cancer then one would expect to observe an overlap between the genes and pathways identified by an EWAS of smoking and GWAS of lung cancer. This specific example has been examined before, with studies suggesting either methylation at two sites (of over 1000 smoking-related sites) mediate over 30% of the effect of smoking on lung cancer⁵⁵ or that there is little evidence for a causal effect of DNAm on lung cancer¹¹. The results of either study suggest most sites will not mediate the effect of smoking on lung cancer and thus there would be little overlap between genes and pathways of an EWAS of smoking and a GWAS of lung cancer. This is corroborated by the results of our study: there was little evidence that correlation of pathway enrichment scores between the two, 0.15, was greater than the mean correlation enrichment score across all GWAS-GWAS and GWAS-EWAS correlations, 0.12 (FDR > 0.05).

It is important to note that overlap between GWAS and EWAS genesets may be missed even if this mediation model is true for various traits. As shown in the simulations, detecting this overlap depends on individual study power as well as the underlying genetic architecture of the trait. There are further things that may limit the detection of geneset overlap that are discussed in the limitations below.

Information gained from EWAS

The fact that genes and pathways identified from some GWAS and EWAS of the same traits are seemingly very separate suggests we are gaining new information from EWAS, even if interpreting the new information may be difficult. Key to interpreting the EWAS results would be to try and disentangle whether the EWAS results are likely due to confounding. Interpreting EWAS can also be difficult due to cell type heterogeneity and the complexity of mechanisms which mediate DNAm changes^{3,8,56–58}. Due to these difficulties, it should not be concluded that EWAS definitely help increase our biological understanding of complex traits. Rather, DNAm is capturing different biological information. Regardless of biological insight gained, the translational impact may still be gleaned from DNAm studies; DNAm may aid diagnoses by acting as a reliable biomarker or could help predict various health outcomes².

There are also benefits to understanding the biological consequences of a trait, something that EWAS might help identify and GWAS will not (at least not directly). This does depend on further research to understand how changes in DNAm downstream to complex trait variation is relevant to human health. Further, establishing where exactly DNAm may lie on the causal pathway may be difficult

and work is ongoing to discover this for various traits²². Use of causal inference methods such as Mendelian randomisation^{27,38,59} have been applied, but this still comes with various caveats^{22,59}. Some studies also try and confirm effects experimentally¹¹ and use previous biological knowledge of the trait to try and understand EWAS results.

However, some of the biological knowledge of complex traits relating to genes and pathways comes from GWAS of those traits. This study suggests that EWAS is unlikely to identify many genes proximal to genetic variation pertinent to the trait of interest and further the genesets are unlikely to overlap with those identified in GWAS. Therefore, comparison of EWAS results to those of a corresponding GWAS is unlikely to yield much insight. This may make inference from EWAS difficult, yet it seems likely the interpretability of DNAm studies will continue to improve over the coming years, as understanding the underlying epigenetic architecture of complex traits could still provide translational benefits².

Limitations

As discussed, detecting gene or pathway overlap depends on the genetic and DNAm architecture of the trait. Here only 15 traits, two of which are smoking behaviour traits and two of which are blood pressure traits, have been studied. Further, these traits are mostly exposures that precede disease. This means the results cannot be generalised to all or even the majority of complex traits or diseases. These analyses could be repeated by setting a less restrictive sample size limit, but it was felt that would make the results less reliable and impossible in many circumstances where too few DNAm sites had been discovered by EWAS. For each of the EWAS used here, and for the majority of high-powered EWAS, the number of DNAm sites measured is small compared to the total number in the genome (less than 5%). Unfortunately, imputation is not available for inferring DNAm levels, and therefore much of the genome covered by the millions of genetic variants that can be measured (and inferred) using genotyping arrays will have been missed by the arrays used for the EWAS in this study. As sample sizes increase and technologies measuring more DNAm sites become more common, it would be interesting to repeat the analysis.

Often in GWAS and EWAS, prioritisation of SNPs and DMPs identified occurs before functional mapping. Prioritisation for both studies may be informed by prior knowledge of the trait, prior understanding of molecular biology, predicted consequences of observed variation (for example Ensembl's Variant Effect Predictor⁶⁰), replication of findings or a number of other methods. In this study, we did not perform any prioritisation (besides the conventional P value threshold cutoffs) and thus may have increased the amount of "noise" in the signal taken forward for functional annotation. Unfortunately, this extra prioritisation of sites is not tractable when comparing many different association studies and may reduce power to detect any overlap between genes and pathways. However, added noise is unlikely to prevent the detection of true overlap if that true overlap is substantial, as shown by our simulations (Fig. 3).

The nearest gene, by chromosomal position, to a DMP or SNP is not necessarily the gene of interest. SNPs may have effects on genes distal to their position⁶¹ and the correlation between genetic variants inflates associations of variants proximal to the true causal variant, which may map to unrelated genes. Further, the correlation structure in DNAm data may induce associations between complex traits at a site far from where variation in DNAm causes complex trait changes⁶¹. Therefore, the mapping of DMPs and SNPs to genes in this study could likely be improved. The "correct" method for this mapping has not yet been established though and even though some tools are available (such as eQTL studies), there are caveats to them too^{62,63}. Further, recent work by Mountjoy et al. suggests that the best predictor for causal gene(s) is the distance between the gene(s) and GWAS signal⁶⁴.

Our understanding of molecular pathways is not complete and thus attributing genes to certain pathways or functionalities may be

erroneous. However, the results remained consistent across four different methods that annotate genes to pathway, suggesting differences in mapping genes to genesets should not impact our conclusions.

Biological information gained from GWAS and EWAS may be defined in various ways and depending on the interpretation of this, one could alter methods used to extract biological information. However, first exploring the genomic regions identified and then mapping these to potentially relevant biological pathways is common amongst GWAS and EWAS^{4-7,65-67} and provided a simple way to compare the information from the two study types.

A simple explanation of EWAS genes not overlapping with GWAS genes is that they are not causal. However, it is possible that GWAS genes are limited to a small subset of causal genes in which functional variation is evolutionarily permissible. Because epigenetic variation can be modified by the environment it's possible that causal genes that are unavailable to GWAS through lack of functional variation are identified through EWAS. Such a scenario would reduce the power to detect overlap using the simulation approaches that we employed here. Another scenario is that the causal nature of DNA methylation is similarly multi-factorial as the genetic component, which would entail that with low power EWAS hits are more likely to be reverse causal, but forward causal EWAS hits (that by necessity have small effects) will only be detected when sample sizes and genomic coverage rival that of GWAS.

Overall, this study provides evidence that, for 13 of 15 complex traits, there is little overlap between genes and genesets identified by GWAS and EWAS of the same trait. Given the differences in properties between DNAm and genetic variants the results presented in this study may apply to other traits, but this is still to be confirmed. Where lack of overlap between genes and genesets is found, it suggests EWAS may be providing new biological information, however, the interpretability of EWAS is still in question and with current methods it is hard to determine if EWAS results are attributable to confounding or reverse causation. Regardless, as datasets grow and causal inference methods improve we are likely to be better able to interpret the role of DNAm in complex traits.

Methods

Samples

EWAS summary data and GWAS summary data were extracted from The EWAS Catalog²³ and the IEU OpenGWAS Project^{31,32} respectively. For traits that had multiple EWAS with a sample size of greater than 4500, the EWAS with the largest sample size was used, the same was applied to the GWAS. The sample size, first authors, and PubMed IDs can be found in Table 1.

All GWAS were conducted in European populations. EWAS were conducted in European populations only, or were part of trans-ancestry meta-analyses, that contained European individuals and heterogeneity analyses were conducted within each study showing high correlation of associations between populations.

Overlapping genomic regions

Each chromosome was divided into 500 Kb blocks, each block that did not contain a DNAm site measured by the Illumina Infinium HumanMethylation450 BeadChip was removed. For each trait, the genome blocks that had one or more EWAS sites and one or more GWAS sites that reached the set *p*-value threshold were tallied. The *p*-value threshold was set at a lenient $P < 1 \times 10^{-5}$ or if it was lower, the maximum reported *p*-value in the EWAS of that trait.

Mapping sites to genes and genesets

The R package biomaRt⁶⁸ was used to extract Ensembl gene ids along with chromosome positions of all genes. The package was also used to extract gene ontology (GO) terms^{24,25} and map these to the Ensembl

gene ids. The R package *limma*⁶⁹ was used to extract KEGG terms^{26–28} and these were mapped to Ensembl gene ids.

Protein–protein interaction data, which includes data from StringDB⁷⁰ and IntAct⁷¹, and terms from the Reactome database²⁹ were extracted from EpiGraphDB³⁰.

CpG sites associated with traits at $P < 1 \times 10^{-7}$ and SNPs associated with traits at $P < 5 \times 10^{-8}$ were taken forward to be mapped to genes. The correlation structure present in genetic and DNAm data makes it difficult to ascertain the precise site driving any signal observed. Thus, no filtering based on correlation between variants or CpG sites was performed.

For each CpG site identified by EWAS and used in the analyses, it was mapped to the nearest gene (Ensembl gene ID) by chromosome position. If a CpG site lay within the bounds of multiple genes then the site was mapped to all of those genes. Therefore, one CpG site could map to multiple genes and one gene could map to multiple CpG sites. The same gene mapping approach was used for variants identified in GWAS. The positions of CpG sites were extracted using the R package *meffil*⁷².

Methods for assessing overlap

To test the overlap between genes identified we generated ORs as so

$$OR_{\text{gene–overlap}} = \frac{\text{odds}_{EG}}{\text{odds}_{EnG}} \quad (1)$$

where odds_{EG} is the odds of a gene being identified in EWAS and GWAS and odds_{EnG} is the odds of a gene being identified in EWAS, but not in GWAS.

Genes may map to genesets by chance. Often in GWAS and EWAS, enrichment for any genesets are tested by assessing whether the genes identified are more common in any geneset than expected by chance. We tested mapping genes to genesets and directly assessing overlap like in Eq. (1) against correlation between enrichment scores for each geneset. Enrichment scores are also odds ratios generated in a similar way to those in Eq. (1):

$$OR_{\text{geneset–enrichment}} = \frac{\text{odds}_{GS}}{\text{odds}_{nGS}} \quad (2)$$

where odds_{GS} is the odds of a gene being annotated to the geneset and odds_{nGS} is the odds of a gene not being annotated to the geneset.

For many genesets, genes annotated to that geneset would not be identified in an GWAS or EWAS, causing the enrichment score for many genesets to be zero. Further, some genesets would have very large enrichment scores. This made the relationship between enrichment scores generated for the EWAS results and the enrichment scores generated for the GWAS non-normal. Thus, to examine the relationship between the two, we generated Spearman's rank correlation coefficients for the logarithm of enrichment scores.

Testing power to detect overlap

Simulations were setup as seen in Box 1.

The simulations iterated over each set of parameters 1000 times. For each iteration, two sets of genes, GWAS genes and EWAS genes, were sampled from the total set of genes. Each iteration assessed gene overlap and geneset overlap between these gene sets using Eq. (1). Equation (2) was used to generate enrichment scores for each gene set and then the correlation between the enrichment scores was assessed.

GWAS genes were only sampled from a set of “causal” genes and a proportion of EWAS genes were sampled from the set of causal genes and the rest from the set of “associated” genes. Receiver operator characteristic (ROC) curves were generated to assess whether it was possible to predict the gene overlap, geneset overlap, and enrichment score correlations for scenarios where the proportion of causal EWAS

genes was greater than zero from the scenario where the proportion of causal EWAS genes was zero. The area under these ROC curves were then calculated in each case.

These simulations were repeated for each geneset and protein–protein interaction database. The protein–protein interaction and Reactome databases all map only to protein coding genes, whereas the GO and KEGG databases map to all Ensembl gene IDs. To compare predictive ability across annotation methods, we excluded Ensembl gene IDs that were not protein coding genes. We also compared the performance of models when mapping to all Ensembl gene IDs and protein coding genes only for GO and KEGG databases. (Supplementary Fig. 7).

From these simulations, the best method to assess geneset overlap, and the best geneset annotation method to assess that overlap and the scenarios (i.e., study power required, proportion of DMPs that need to be causal) in which we expect to be able to detect overlap could be deduced.

Empirical analyses

The SNPs identified in the GWAS at $P < 5 \times 10^{-8}$ and the DNAm sites identified in the EWAS at $P < 1 \times 10^{-7}$ and were mapped to genes and genesets. Overlap between genes was calculated as before (Eq. (1)), enrichment scores were generated and correlated as described above.

Expected overlap was generated to compare to the observed results. For this, random positions were chosen in the genome equal to the number of DMPs identified in the EWAS. These genes were then used to assess gene and geneset overlap as for the observed results. This was repeated 1000 times to generate a null distribution and a z-test was used to assess whether there was a difference between the observed results and the mean of the null distribution.

There is a correlation structure within DNAm data⁷³, we hypothesised this might contribute to the observed results. By randomly sampling positions from the genome, a new correlation structure between DMPs would be generated. We tested whether sampling the genome in a non-random way, aimed at keeping some correlation structure, altered the results. To generate new data whilst attempting to keep a similar correlation structure, a fixed number of base pairs were added to each of the DMPs identified in the empirical analysis such that

$$BP_{\text{new}} = BP_{\text{dmp}} + \max(G) \times I \quad (3)$$

where BP_{new} = base pair of new site, BP_{dmp} = base pair of DMP identified in the EWAS, G = gene size, and I = iteration.

If BP_{new} extended beyond the end of a chromosome the position moved onto the next chromosome, with positions moving past the end of chromosome 22, being moved to chromosome 1.

Overall, the overlap between genes and genesets identified by GWAS and EWAS did not change across null distribution sampling methods.

Understanding architecture from geneset overlap

Simulations were setup as illustrated in Supplementary Fig. 4. Here we describe simulations for a single trait. These were repeated for all traits. Firstly, SNPs identified in the GWAS and DMPs identified in the EWAS were mapped to genes as described above in ‘Empirical analyses’. Genes were then randomly sampled from Ensembl gene IDs and were assigned as either “causal,” meaning changes in that gene effect variation in the phenotype across individuals, or “associated,” meaning changes in the gene are associated with the phenotype across individuals, but the nature of association is not known. The empirically identified (or “known”) GWAS genes (KGG) were added to the list of causal genes and the empirically identified (or “known”) EWAS genes (KEG) were added to the list of associated genes. These combined set of causal and associated genes can be thought of as all the genes

BOX 1**Box 1: Simulations to determine the power of assessing overlap between GWAS and EWAS genes/genesets**

<u>Steps</u>			
1. Set parameters 2. Generate set of GWAS genes and generate a set of EWAS genes 3. Map genes to genesets and generate enrichment scores 4. Assess overlap of genes, overlap of genesets, and correlation of geneset enrichment scores			
Step 1	<table border="0"> <tr> <td style="vertical-align: top;"> <u>Variable parameters</u> - Number of causal genes (NG) - EWAS power (EP) - Proportion of EWAS genes that are causal (PEC) - Geneset database </td> <td style="vertical-align: top;"> <u>Static parameters</u> - GWAS power (GP) = 0.1 - Total genes (TG) = all Ensembl gene IDs = 60,575 </td> </tr> </table>	<u>Variable parameters</u> - Number of causal genes (NG) - EWAS power (EP) - Proportion of EWAS genes that are causal (PEC) - Geneset database	<u>Static parameters</u> - GWAS power (GP) = 0.1 - Total genes (TG) = all Ensembl gene IDs = 60,575
<u>Variable parameters</u> - Number of causal genes (NG) - EWAS power (EP) - Proportion of EWAS genes that are causal (PEC) - Geneset database	<u>Static parameters</u> - GWAS power (GP) = 0.1 - Total genes (TG) = all Ensembl gene IDs = 60,575		
Step 2	<table border="0"> <tr> <td style="vertical-align: top;"> <u>Generate causal gene list</u> - Causal genes (CaG): sample NG genes from TG <u>Generate GWAS gene list</u> - N GWAS genes = NG * GP - To ascertain GWAS genes: sample N GWAS genes from CaG </td> <td style="vertical-align: top;"> <u>Generate EWAS gene list</u> - N EWAS genes = NG * EP - EWAS genes will be sampled from the causal genes AND from TG (made up of both causal and non-causal genes) - To ascertain causal EWAS genes: sample N EWAS genes * PEC from CaG - To ascertain non-causal EWAS genes: sample N EWAS genes * (1-PEC) from TG - Combine non-causal and causal EWAS genes </td> </tr> </table>	<u>Generate causal gene list</u> - Causal genes (CaG): sample NG genes from TG <u>Generate GWAS gene list</u> - N GWAS genes = NG * GP - To ascertain GWAS genes: sample N GWAS genes from CaG	<u>Generate EWAS gene list</u> - N EWAS genes = NG * EP - EWAS genes will be sampled from the causal genes AND from TG (made up of both causal and non-causal genes) - To ascertain causal EWAS genes: sample N EWAS genes * PEC from CaG - To ascertain non-causal EWAS genes: sample N EWAS genes * (1-PEC) from TG - Combine non-causal and causal EWAS genes
<u>Generate causal gene list</u> - Causal genes (CaG): sample NG genes from TG <u>Generate GWAS gene list</u> - N GWAS genes = NG * GP - To ascertain GWAS genes: sample N GWAS genes from CaG	<u>Generate EWAS gene list</u> - N EWAS genes = NG * EP - EWAS genes will be sampled from the causal genes AND from TG (made up of both causal and non-causal genes) - To ascertain causal EWAS genes: sample N EWAS genes * PEC from CaG - To ascertain non-causal EWAS genes: sample N EWAS genes * (1-PEC) from TG - Combine non-causal and causal EWAS genes		
Step 3	<table border="0"> <tr> <td style="vertical-align: top;"> <u>Map genes to genesets</u> - Map all EWAS genes to genesets - Map all GWAS genes to genesets </td> <td style="vertical-align: top;"> <u>Generate geneset enrichment scores</u> - Calculate the enrichment of GWAS genes in all genesets to generate GWAS enrichment scores - Calculate the enrichment of EWAS genes in all genesets to generate EWAS enrichment scores </td> </tr> </table>	<u>Map genes to genesets</u> - Map all EWAS genes to genesets - Map all GWAS genes to genesets	<u>Generate geneset enrichment scores</u> - Calculate the enrichment of GWAS genes in all genesets to generate GWAS enrichment scores - Calculate the enrichment of EWAS genes in all genesets to generate EWAS enrichment scores
<u>Map genes to genesets</u> - Map all EWAS genes to genesets - Map all GWAS genes to genesets	<u>Generate geneset enrichment scores</u> - Calculate the enrichment of GWAS genes in all genesets to generate GWAS enrichment scores - Calculate the enrichment of EWAS genes in all genesets to generate EWAS enrichment scores		
Step 4	<u>Assess overlap</u> - Assess enrichment of EWAS genes in GWAS genes - Assess enrichment of EWAS genesets in GWAS genesets - Generate Pearson's correlation coefficient between GWAS enrichment scores and EWAS enrichment scores		

+Note this may sample genes from CaG, because genes can be both on the causal pathway to a trait and downstream from a trait so EWAS might identify causal genes even when the DNAmethylation variation detected was downstream to trait variation

related to the trait of interest. A number of genes, equal to the number of KGG (N_{KGG}), was sampled from the causal set of genes and assigned to be the “GWAS genes” in the simulations. A number of genes, equal to the number of KEG (N_{KEG}), was sampled from the associated set of genes and assigned to be the “EWAS genes” in the simulations. Then geneset enrichment analyses for both the GWAS and EWAS genes were performed (Eq. (2)) and correlation between the enrichment scores was assessed as previously. In these simulations, the number of total genes was varied and the number of causal and associated genes was always set to be half of the total number of genes related to a trait. The total number of genes was proportional to the total number of known genes ($N_{KTG} = N_{KGG} + N_{KEG}$). In each simulation, the number of associated genes (and causal genes) equalled N_{KTG} discovered multiplied by 1, 2, 3, 5, 10, or 20. Therefore, the smallest number of total genes for any simulation was double the number of N_{KTG} and the greatest number of total genes was 40 times N_{KTG} . The other variable set to vary between simulations was the proportion of overlap between causal and associated genes. The proportion of overlap was 0, 0.01, 0.1, 0.5, or 1, where 0 represented the scenario where only the overlap in KGG and KEG would be present in the overlap between causal and associated genes and 1 represented the scenario where the only “non-causal” genes would be the KEG that did not overlap with KGG. For

each simulation scenario, the simulations were repeated 1000 times and box plots show the range of output from those 1000 repeats. Evidence for a difference in the empirically determined correlation of geneset enrichment scores and the mean correlation of geneset enrichment scores across simulations was assessed using a z-test for difference.

Assessing the correlation between geneset enrichment results
 GWAS were extracted from the IEU OpenGWAS Project^{31,32} with the following criteria:

- Sample size > 5000
- European population
- For binary traits, number of cases and controls had to be greater than 500
- Full genome-wide results, i.e., not just associations between a molecular trait and variants in cis.

For each GWAS, all SNPs that associated with the trait at $P < 5 \times 10^{-8}$ were extracted. CpGs associated with the 15 EWAS at $P < 1 \times 10^{-7}$ were then extracted and mapped to genes. For each study, enrichment scores were generated for GO terms as before (Eq. (2)) and correlation between them assessed.

When assessing whether gene overlap or geneset enrichment score correlations were greater than the mean, a z-test was performed. As multiple tests were performed we applied the Benjamini–Hochberg method⁷⁴ to limit the false discovery rate.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the data used in this analysis is publicly available. Thirteen EWAS summary statistics can be downloaded from The EWAS Catalog²³(<http://ewascatalog.org/>) under accession codes: 28002404_body_mass_index_discovery_and_replication, 27651444_smoking_current_vs_never_smoking, 27651444_smoking_former_vs_never_smoking, 27843151_alcohol_consumption_per_day_european_ancestry, 27955697_creative_protein_discovery, 29086770_educational_attainment_basic_model, 31197173_Liu-J_fasting_glucose_base, 31197173_Liu-J_fasting_insulin_base, 29198723_Richard-MA_systolic_blood_pressure_meta-analysis, 29198723_Richard-MA_diastolic_blood_pressure_meta-analysis, 31015461_Kupers-L_birthweight_meta_analysis_all_ancestries, 29311653_Marioni-R_cognitive_abilities_digit_test_basic_adjusted_model, 31073081_Imboden-M_fev1_meta-analysis. The two EWAS not included here are for eGFR and urate. The eGFR EWAS summary statistics was downloaded from <https://ckdgen.imbi.uni-freiburg.de/files/Schlosser2021/eGFR.csv.zip> and the urate summary statistics was downloaded from <https://ckdgen.imbi.uni-freiburg.de/files/Tin2021/urate.csv.zip>. The GWAS summary statistics can be downloaded from the OpenGWAS Project^(21,32)(<https://gwas.mrcieu.ac.uk/>) under accession codes: ieu-b-40, ieu-b-4877, ukb-b-2134, ieu-b-73, ieu-b-35, ieu-a-1239, ebi-a-GCST005186, ebi-a-GCST005185, ieu-b-38, ieu-b-39, ieu-a-1083, ebi-a-GCST006572, ukb-b-19657. The eGFR GWAS summary statistics were downloaded from https://ckdgen.imbi.uni-freiburg.de/files/Stanzick2021/metal_eGFR_meta_ea1.TBL.map.annot.gc.gz and the urate GWAS summary statistics were downloaded from https://ckdgen.imbi.uni-freiburg.de/files/Tin2019/urate_chr1_22_LQ_IQ06_mac10_EA_60_prec1_nstud30_summac400_rsid.txt.gz. The data can also be obtained from the original papers, the PubMed IDs can be found in Table 1. The FEV1 and former versus never smoking GWAS summary statistics are not published and were conducted within the MRC Integrative Epidemiology Unit and uploaded to the OpenGWAS Project database. Their GWAS IDs within the OpenGWAS Project are ukb-b-19657 and ukb-b-2134 respectively. v6.5.6 of the OpenGWAS Project was used to download the data. The R package biomaRt (v2.50.3)⁶⁸ was used to extract Ensembl gene ids along with chromosome positions of all genes. The package was also used to extract gene ontology (GO) terms^{24,25} and map these to the Ensembl gene ids. The R package limma (v3.50.3)⁶⁹ was used to extract KEGG terms^{26–28} and these were mapped to Ensembl gene ids. Protein–protein interaction data, which includes data from StringDB⁷⁰ and IntAct⁷¹, and terms from the Reactome database²⁹ were extracted from EpiGraphDB (v1.0)³⁰.

Code availability

Code used to run the analyses is available here: <https://zenodo.org/badge/latestdoi/521364771>. The code is also on GitHub: <https://github.com/thomasbatram/ewas-gwas-comparisons>. Analyses were completed using R (version 3.6.2) and Python (version 3.7.4).

References

- Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12**, 529–541 (2011).
- Relton, C. L. & Davey Smith, G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS Med.* **7**, e1000356 (2010).
- Birney, E., Smith, G. D. & Grealley, J. M. Epigenome-wide association studies and the interpretation of disease-omics. *PLoS Genet.* **12**, e1006105 (2016).
- Sharp, G. C. et al. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium. *Human Mol. Genet.* **26**, 4067–4085 (2017).
- Reese, S. E. et al. Epigenome-wide meta-analysis of DNA methylation and childhood asthma. *J. Allergy Clin. Immunol.* **143**, 2062–2074 (2019).
- Everson, T. M. et al. Epigenome-wide analysis identifies genes and pathways linked to neurobehavioral variation in preterm infants. *Sci. Rep.* **9**, 6322 (2019).
- Chen, J. et al. Association of a reproducible epigenetic risk profile for schizophrenia with brain methylation and function. *JAMA Psychiatry* **77**, 628–636 (2020).
- Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
- Illingworth, R. S. & Bird, A. P. CpG islands - 'A rough guide' *FEBS letters* **11**, 1713–1720 (2009).
- Breeze, C. E. et al. eFORGE: A tool for identifying Cell type-specific signal in epigenomic data. *Cell Rep.* **17**, 2137–2150 (2016).
- Batram, T. et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. *Int. J. Epidemiol.* **48**, 1493–1504 (2019).
- Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 7171–730 (2018).
- Margolis, S. S., Sell, G. L., Zbinden, M. A. & Bird, L. M. Angelman syndrome. *Neurotherapeutics* **12**, 641–650 (2015).
- Cassidy, S. B., Schwartz, S., Miller, J. L. & Driscoll, D. J. Prader-Willi syndrome. *Genet. Med.* **14**, 10–26 (2012).
- Relton, C. L. & Davey Smith, G. Mendelian randomization: applications and limitations in epigenetic studies. *Epigenomics* **7**, 1239–1243 (2015).
- Min, J. L. et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet.* **53**, 1311–1321 (2021).
- Wahl, S. et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
- Morris, T. T., Davies, N. M., Hemani, G. & Smith, G. D. Population phenomena inflate genetic associations of complex social traits. *Sci. Adv.* **6**, eaay0328 (2020).
- Brumpton, B. et al. Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nat. Commun.* **11**, 1–13 (2020).
- Howe, L. J. et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat. Genet.* <https://doi.org/10.1038/s41588-022-01062-7> (2022).
- Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
- Neumeyer, S., Hemani, G. & Zeggini, E. Strengthening causal inference for complex disease using molecular quantitative trait loci. *Trends Mol. Med.* **26**, 232–241 (2020).
- Batram, T. et al. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res.* **7**, 41 (2022).
- Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
- Carbon, S. et al. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
- Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic. acids res.* **28**, 27–30 (2000).

27. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–D595 (2019).
28. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein sci.* **28**, 1947–1951 (2019).
29. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
30. Liu, Y. et al. Epigraphdb: a database and data mining platform for health data science. *Bioinformatics* **37**, 1304–1311 (2021).
31. Elsworth, B. et al. The MRC IEU OpenGWAS data infrastructure. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.08.10.244293> (2020).
32. Hemani, G. et al. The MR-base platform supports systematic causal inference across the human genome. *eLife* **7**, e34408 (2018).
33. Kurushima, Y. et al. Epigenetic findings in periodontitis in UK twins: A cross-sectional study. *Clin. Epigenet.* **11**, 27 (2019).
34. Agha, G. et al. Blood leukocyte DNA methylation predicts risk of future myocardial infarction and coronary heart disease. *Circulation* **140**, 645–657 (2019).
35. Odintsova, V. V. et al. DNA methylation signatures of breastfeeding in buccal cells collected in mid-childhood. *Nutrients* **11**, 2804 (2019).
36. Geurtsen, M. L., Jaddoe, V. W., Salas, L. A., Santos, S. & Felix, J. F. Newborn and childhood differential DNA methylation and liver fat in school-age children. *Clin. Epigenet.* **12**, 3 (2019).
37. Davey Smith, G. & Ebrahim, S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, (2003).
38. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–98 (2014).
39. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
40. van Dongen, J. et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.* **7**, 11115 (2016).
41. Hannon, E. et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet.* **14**, e1007544 (2018).
42. Lawlor, D. A., Smith, G. D., Kundu, D., Bruckdorfer, K. R. & Ebrahim, S. Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? *Lancet* **363**, 1724–1727 (2004).
43. Sattar, N. & Preiss, D. Reverse causality in cardiovascular epidemiological research: more common than imagined? *Circulation* **24**, 2369–2372 (2017).
44. Davey Smith, G. & Phillips, A. N. Confounding in epidemiological studies: why ‘independent’ effects may not be all they seem *BMJ* **305**, 757–759 (1992).
45. Fewell, Z., Davey Smith, G. & Sterne, J. A. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am. J. Epidemiol.* **166**, 646–655 (2007).
46. Heijmans, B. T. & Mill, J. Commentary: the seven plagues of epigenetic epidemiology. *Int. J. Epidemiol.* **41**, 74–78 (2012).
47. Shah, S. et al. Improving phenotypic prediction by combining genetic and epigenetic associations. *Am. J. Human Genet.* **97**, 75–85 (2015).
48. Trejo Banos, D. et al. Bayesian reassessment of the epigenetic architecture of complex traits. *Nat. Commun.* **11**, 2865 (2020).
49. Philibert, R. et al. Reversion of AHRR demethylation is a quantitative biomarker of smoking cessation. *Front. Psychiatry* **7**, 55 (2016).
50. Richard, M. A. et al. DNA methylation analysis identifies loci for blood pressure regulation. *Am. J. Human Genet.* **101**, 888–902 (2017).
51. Kato, N. et al. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* **47**, 1282–1293 (2015).
52. Tin, A. et al. Epigenome-wide association study of serum urate reveals insights into urate co-regulation and the *slc2a9* locus. *Nat. Commun.* **12**, 1–18 (2021).
53. Wheeler, H. E. et al. Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS Genet.* **12**, e1006423 (2016).
54. Mogil, L. S. et al. Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* <https://doi.org/10.1371/journal.pgen.1007586> (2018).
55. Fasanelli, F. et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* **6**, 10192 (2015).
56. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform.* **13**, 86 (2012).
57. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
58. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).
59. Relton, C. L. & Davey Smith, G. Two-step epigenetic mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int. J. Epidemiol.* **41**, 161–176 (2012).
60. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
61. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
62. Huang, Q. Q., Ritchie, S. C., Brozynska, M. & Inouye, M. Power, false discovery rate and Winner’s Curse in eQTL studies. *Nucleic Acids Res.* **46**, e133 (2018).
63. Kanduri, C., Bock, C., Gundersen, S., Hovig, E. & Sandve, G. K. Colocalization analyses of genomic elements: approaches, recommendations, and challenges. *Bioinformatics* **35**, 1615–1624 (2019).
64. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
65. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in 700,000 individuals of European ancestry. *Human Mol. Genet.* **27**, 3641–3649 (2018).
66. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation *Am J Hum Genet.* **101**, 5–22 (2017).
67. Ellinghaus, D. et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–8 (2016).
68. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–91 (2009).
69. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
70. Szklarczyk, D. et al. STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
71. Orchard, S. et al. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).
72. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).

73. Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T. & Engelhardt, B. E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.* **16**, 14 (2015).
74. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodological)* **57**, 289–300 (1995).
75. Joehanes, R. et al. Epigenetic signatures of cigarette smoking. *Circ.: Cardiovasc. Genet.* **9**, 436–447 (2016).
76. Liu, M. et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
77. Liu, C. et al. A DNA methylation biomarker of alcohol consumption. *Mol. Psychiatry* **23**, 422–433 (2016).
78. Ligthart, S. et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol.* <https://pubmed.ncbi.nlm.nih.gov/27955697/> (2016).
79. Ligthart, S. et al. Genome analyses of >200,000 individuals identify 58 loci for chronic inflammation and highlight pathways that link inflammation and complex disorders. *Am. J. Hum. Genet.* **103**, 691–706 (2018).
80. Karlsson Linnér, R. et al. An epigenome-wide association study meta-analysis of educational attainment. *Mol. Psychiatry* **22**, 1680–1690 (2017).
81. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
82. Manning, A. K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
83. Evangelou, E. et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* **50**, 1412–1425 (2018).
84. Küpers, L. K. et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. *Nat. Commun.* **10**, 1–11 (2019).
85. Horikoshi, M. et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).
86. Marioni, R. E. et al. Meta-analysis of epigenome-wide association studies of cognitive abilities. *Mol. Psychiatry* **23**, 2133–2144 (2018).
87. Imboden, M. et al. Epigenome-wide association study of lung function level and its change. *Eur. Respir. J.* <https://pubmed.ncbi.nlm.nih.gov/31073081/> (2019).
88. Schlosser, P. et al. Meta-analyses identify DNA methylation associated with kidney function and damage. *Nat. Commun.* **12**, 1–16 (2021).
89. Stanzick, K. J. et al. Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat. Commun.* **12**, 1–17 (2021).
90. Tin, A. et al. Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat. Genet.* **51**, 1459–1474 (2019).

Acknowledgements

This work was supported by the Wellcome Trust through a Wellcome PhD studentship to T.B. [203746]. G.H. is funded by the Wellcome Trust and Royal Society [208806/Z/17/Z]. This work was carried out in the MRC Integrative Epidemiology Unit, which is supported by the UK Medical Research Council (MC_UU_00011/1 and MC_UU_00011/4). We are grateful to Dr. Kaitlin Wade for highly valuable comments.

Author contributions

T.B. and G.H. conceived the ideas and designed the study. T.B. conducted the main analyses. G.H. conducted analyses present in the supplementary note found within Supplementary Information. T.B. wrote the manuscript. T.R.G., C.L.R., and N.J.T. contributed valuable study design ideas and manuscript edits. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-35037-3>.

Correspondence and requests for materials should be addressed to Thomas Battram.

Peer review information *Nature Communications* thanks Norihiro Kato and Pascal Schlosser for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022