Disease variants alter transcription factor levels and methylation of their binding sites

Nature genetics, doi:10.1038/ng.3721

Marc Jan Bonder1*, René Luijk2*, Daria V. Zhernakova1, Matthijs Moed2, Patrick Deelen1,2, Martijn Vermaat4, Maarten van Iterson2, Freerk van Dijk1,3, Michiel van Galen3, Jan Bot5, Roderick C. Slieker2, P. Mila Jhamai6, Michael Verbiest8, H. Eka D. Suchiman2, Marijn Verkerk6, Ruud van der Breggen2, Jeroen van Rooij6, Nico Lakenberg2, Wibowo Arindrarto8, Szymon M. Kielbasa7, Iris Jonkers2, Peter van 't Hof6, Irene Nooren5, Marian Beekman2, Joris Deelen2, Diana van Heemst9, Alexandra Zhernakova1, Ettje F. Tigchelaar1, Morris A. Swertz1,3, Albert Hofman10, André G. Uitterlinden9, René Pool11, Jenny van Dongen11, Jouke J. Hottenga11, Coen D.A. Stehouwer12,13, Carla J.H. van der Kallen12,13, Casper G. Schalkwijk12,13, Leonard H. van den Berg14, Erik. W. van Zwet6, Hailiang Mei7, Yang Li1, Mathieu Lemire15, Thomas J. Hudson15,16,17, BIOS Consortium18, P. Eline Slagboom2, Cisca Wijmenga1, Jan H. Veldink14, Marleen M.J. van Greevenbroek12,13, Cornelia M. van Duijn19, Dorret I. Boomsma11, Aaron Isaacs19, Rick Jansen20, Joyce B.J. van Meurs6, Peter A.C. ’t Hoen4,#, Lude Franke1,#, Bastiaan T. Heijmans2,#
Abstract

Most disease-associated genetic variants are non-coding, making it challenging to design experiments to understand their functional consequences\textsuperscript{1,2}. Identification of expression quantitative trait loci (eQTLs) has been a powerful approach to infer downstream effects of disease variants but the large majority remains unexplained\textsuperscript{3,4}. The analysis of DNA methylation, a key component of the epigenome\textsuperscript{5,6}, offers highly complementary data on the regulatory potential of genomic regions\textsuperscript{7,8}. Here, we show that disease variants have widespread effects on DNA methylation in trans that likely reflect differential occupancy of trans-binding sites by cis-regulated transcription factors. Using multiple omics data on 3,841 Dutch individuals, we identified 1,907 established trait-associated SNPs that affect methylation levels of 10,141 different CpG sites in trans (FDR<0.05). These included SNPs that affect both the expression of a nearby transcription factor (like NFKB1, CTCF and NKX2-3) and methylation of its respective binding site across the genome. Trans-meQTLs effectively expose downstream effects of disease-associated variants.

1. Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands; 2. Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; 3. Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands; 4. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; 5. SURFsara, Amsterdam, the Netherlands; 6. Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands; 7. Sequence Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands; 8. Medical Statistics Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; 9. Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands; 10. Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands; 11. Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands; 12. Department of Internal Medicine, Maastricht University Medical Center, Maastricht, The Netherlands; 13. School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands; 14. Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands; 15. Ontario Institute for Cancer Research, Toronto, Ontario, Canada M5G 0A3; 16. Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada M5S 1A1; 17. Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A1; 18. A full list of members and affiliations appears in the Supplementary Note; 19. Genetic Epidemiology Unit, Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands; 20. Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands; * These authors contributed equally to this work; # These authors jointly directed this work; Correspondance to: Lude Franke (lude@ludesign.nl) and Bastiaan T. Heijmans (bas.heijmans@lumc.nl)
To systematically study the role of DNA methylation in explaining downstream effects of genetic variation, we analysed genome-wide genotype and DNA methylation in whole blood from 3,841 samples from five Dutch biobanks (Figure 1, Supplementary Table 1, Supplemental Text). We found cis-meQTL effects for 34.4% of all 405,709 tested CpGs (n=139,566 at a CpG-level FDR of 5%, \( P \leq 1.38 \times 10^{-4} \)), typically with a short physical distance between the SNP and CpG (median distance 10 kb, Supplementary Fig. 1). By regressing out primary meQTLs effect for each of these CpGs and repeating the cis-meQTL mapping, we observed up to 16 independent cis-meQTLs for these CpGs (Supplementary Table 2) totalling 272,037 independent cis-meQTL effects. Few factors determine whether a CpG site shows a cis-meQTL effect except the variance in methylation level of the CpG site involved (Supplementary Fig. 2, Supplementary Fig. 3a). The proportion of methylation variance explained by SNPs, however, is typically small (Supplementary Fig. 3b). When accounting for this strong effect of CpG variation, we find only modest enrichments and depletions for cis-meQTL CpG sites for CpG island and genic annotation (Supplementary Fig. 3e) or when using annotations of biological function based on chromatin segmentations of 27 blood cell types (Figure 2a).

Figure 1. Overview of a genomic region around TMEM176B and characteristics of CpGs associated to meQTLs and eQTM. In the illustration, the relations between a SNP, DNA methylation at nearby CpGs, and the associations with the gene itself are shown. Boxes show the median, the inter-quartile range (IQR). Whiskers show the outer quartile plus 1.5 times the IQR. The top left plot shows the observed methylation Quantitative Trait Locus (meQTL) between cg23533927 and rs7806458. The top right plot shows the observed expression Quantitative Trait Loci (eQTL) between TMEM176B and rs7806458. The observed methylation-expression association (eQTM) between TMEM176B and cg23533927, is shown below the gene. The left part shows the data before correction for the cis-eQTL and cis-meQTL, the eQTM effect after correction for cis-eQTLs and cis-meQTLs is shown on the right. b, Two overlaid pie charts. The inner chart indicates the proportion of tested CpGs harboring meQTLs. Over 35% of all tested CpGs show evidence for harboring a meQTL, either in cis or in trans. The outer chart indicates what CpGs are associated with gene expression in cis (in total 3.2%). c, Replication of peripheral blood trans-meQTLs in lymphocytes.
Figure 2. Characterization of identified cis- and trans-meQTL and eQTM-effects. a-c, Over- or underrepresentation of CpGs for predicted chromatin states for cis-meQTLs, trans-meQTLs and eQTMs. Grey bars reflect uncorrected enrichments, colored bars reflect enrichments after correction for factors influencing the likelihood of harboring a meQTL or eQTM, including methylation variability. Bar graphs show odds ratios and error bars (95% confidence interval). CGI: CpG island; TssA: Active TSS; TssAFlnk: Flanking active TSS; TxFlnk, Transcribed at gene 5’ and 3’; Tx: Strong transcription; TxWk: Weak transcription; EnhG: Genic enhancer; Enh: Enhancer; ZNF/Rpts: ZNF genes and repeats; Het: Heterochromatin; TssBiv: Bivalent/Poised TSS; BivFlnk: Flanking bivalent TSS/Enhancer; EnhBiv: Bivalent enhancer. d, Decision tree for predicting the effect direction of eQTMs. Each subplot shows the distributions for positive (blue) and negative (red) associations for that subset of the data. Dashed vertical lines indicate the optimal split used by the algorithm. The boxes in the leaves indicate the number of positive and negative effects in each of the leaves. e, Receiver operator characteristic curve showing the performance of the decision tree.
We contrasted these modest functional enrichments to CpGs whose methylation levels correlates with gene expression in cis (i.e. mapping expression quantitative trait methylations (eQTMs)), by generating RNA-seq data for 2,101 out of 3,841 individuals in our study. Using a conservative approach that maximally accounts for potential biases (see Methods), we identified 12,809 unique CpGs that correlated to 3,842 unique genes in cis (CpG-level FDR < 0.05). eQTMs were enriched for mapping in active regions, e.g. in and around active transcription start sites (TSSs) (3-fold enrichment, \(P=1.8 \times 10^{-91}\)) and enhancers (2-fold enrichment, \(P=1.1 \times 10^{-139}\), Figure 2b). The majority of eQTMs showed the canonical negative correlation with transcriptional activity (69.2%) but a substantial minority of correlations was positive (30.8%) in line with recent evidence that DNA methylation does not always negatively correlate with gene expression \(^{14}\). As expected, negatively correlated eQTMs were enriched in active regions like active TSSs (3.7-fold enrichment, \(P=9.5 \times 10^{-202}\)). Positive correlations primarily occurred in repressed regions (e.g. Polycomb repressed, 3.4-fold enrichment, \(P=5.8 \times 10^{-103}\)) (Supplementary Fig. 4). The sharp contrast between positively and negatively associated eQTMs enabled us to predict the direction of the correlation. A decision tree trained on the strongest eQTMs (those with an FDR < 9.7 \(\times 10^{-6}\), \(n=5,137\)) using data on histone marks and distance relative to gene, could predict the direction with an area under the curve of 0.83 (95% confidence interval, 0.78-0.87) (Figure 2d, e).

We next ascertained whether trans-meQTLs are biologically informative, since previous trans-eQTL mapping studies demonstrated that identifying trans-expression effects provide a powerful tool to uncover and understand downstream biological effects of disease-SNPs\(^ {3,15,16}\). We focussed on 6,111 SNPs that were previously associated with complex traits and diseases ('trait-associated SNPs', see Methods and Supplementary Table 3). We observed that one-third of these trait-associated SNPs (1,907 SNPs, 31.2%) affect methylation in trans at 10,141 CpG sites, totalling 27,816 SNP-CpG combinations (FDR<0.05, \(P<2.6 \times 10^{-7}\), Figure 3a). This represents a 5-fold increase in the number of CpG sites affected as compared with a previous trans-meQTL mapping study\(^ {17}\). We evaluated whether the GWAS SNP themselves were likely underlying the trans-effects or that the associations could be attributed to another SNP in moderate LD. Of the 1,907 GWAS SNPs with trans-effects, 1,538 (87.2%) were in strong LD with the top SNP (\(R^2 > 0.8\)), indicating that the GWAS SNPs indeed are the driving force behind many of the trans-meQTLs. Of note, due to the sparse coverage of the Illumina 450k array, the true number of CpGs in the genome that are altered by these trait associated SNPs will be substantially higher.

To validate our trans-meQTLs, we performed a replication analysis in a set of 1,748 lymphocyte samples\(^ {17}\). Of the 18,764 overlapping trans-meQTLs, 94.9% had a consistent allelic direction (Figure 1E; Supplementary Table 4). This indicates that the identified trans-meQTLs are robust and not caused by differences in cell-type composition. Further analysis of SNPs known to influence blood cell composition\(^ {18,19}\) showed no or only few trans-effects and alternative adjustments of the methylation-data corroborated the stability of trans-effects, both indicating a limited influence of cell type composition (Supplementary Results, Supplementary Tables 5–7).

After the identification of the trans-meQTLs, we assessed if the trans-SNPs also affected expression of the genes associated with the trans-CpGs. By overlaying the trans-meQTLs and cis-eQTMs, we could link 436 SNPs to 850 genes, totalling 2,889 SNP-gene pairs. We found significant associations (trans-eQTLs) (FDR < 0.05) for 8.4% of these effects, and 91% of these effects showed the expected direction of the effect, given the directions of the trans-meQTLs and cis-eQTMs (Supplementary Table 8).
In contrast to cis-meQTL CpGs, trans-meQTLs CpGs show substantial functional enrichments: they are enriched around TSSs and depleted in heterochromatin (Figure 2c) and are strongly enriched for being an eQTM (1,913 CpGs (18.9%), 5.2-fold, \(P = 2.3 \times 10^{-101}\)). Among the 1,907 trait-associated SNPs that make up the trans-meQTLs there was an overrepresentation of GWAS-identified SNPs associated with immune- and cancer-related traits (Figure 3a). The large majority of trans-meQTLs were inter-chromosomal (93%, 9,429 CpG-SNP pairs) and included 12 trans-meQTL SNPs (yielding 3,616 unique CpG-SNP pairs) that each showed downstream trans-meQTL effects across all of the 22 autosomal chromosomes (i.e. trans-bands, Figure 3b).

We subsequently studied the nature of these trans-meQTLs. Using high-resolution Hi-C data\(^\text{20}\), we identified 720 SNP-CpG pairs (including 402 CpG sites and 172 SNPs) among the trans-meQTLs that overlapped with an inter-chromosomal contact, which is 2.9-fold more frequent than expected by chance \(P = 3.7 \times 10^{-126}\), Figure 3). The enrichment for Hi-C inter-chromosomal contacts remained after removing SNPs that were responsible for trans-bands \(P = 1.7 \times 10^{-61}\). Hence, inter-chromosomal contacts may produce associations between SNPs and CpGs in trans. In order to characterize the 720 SNP-CpG pairs overlapping with inter-chromosomal contacts, we performed motif enrichments using three motif enrichment analyses (Homer, PWMEnrich, DEEPbind)\(^\text{21–23}\). These analyses revealed that the 402 CpG sites involved frequently overlapped with CTCF, RAD21 and SMC3 binding sites \(P = 2.3 \times 10^{-5}\), \(P = 3.5 \times 10^{-5}\) and \(P = 5.1 \times 10^{-5}\), respectively), factors known to regulate chromatin architecture\(^\text{24,25}\). An analysis of ChIP-Seq data on CTCF binding confirmed this finding (1.8-fold enrichment, \(P = 5.2 \times 10^{-7}\)).
We next tested whether the \textit{trans}-meQTLs reflected the effect of differential transcription factor (TF) binding of TFs that map close to the SNPs. The rationale for this hypothesis is that binding of TFs has been linked to changes in local DNA methylation, primarily loss-of-methylation upon TF binding and gain-of-methylation after loss of TF occupancy\cite{7,8}. This model suggests that \textit{trans}-meQTLs may be attributed to SNPs affecting the expression of a TF \textit{in cis} and that the SNP allele preferentially has a unidirectional effect on DNA methylation. In line with this prediction, we observed that if a SNP is associated with multiple CpG sites \textit{in trans} (at least 10, \(n=305\)), the direction of the association of the SNP was consistently skewed towards either increased or decreased DNA methylation. On average 76\% of the CpGs per \textit{trans}-meQTL SNP displayed the same direction of effect (expected 50\%, \(P=10^{-111}\); Figure 4a).

A significant skew in direction of the allelic effect was present for 59.7\% of the 305 individual SNPs with at least 10 \textit{trans}-meQTL effects and increased to 95.2\% for the 104 SNPs with at least 50 \textit{trans}-meQTL effects (binomial test \(P<0.05\)), suggesting that differential TF binding may explain a substantial fraction of \textit{trans}-meQTLs.

In order to explore this mechanism further, we combined ChIP-seq data on TF binding at CpGs and \textit{cis}-expression effects of SNPs to directly examine the involvement of TFs in mediating \textit{trans}-meQTLs. Among trait-associated SNPs influencing at least 10 CpGs \textit{in trans} (\(n=305\)), we identified 13 \textit{trans}-meQTL SNPs with strong support for a role of TFs (Figure 4a).

The most striking example was a locus on chromosome 4 (Figure 4b), where two SNPs (rs3774937 and rs3774959, in strong LD) were associated with ulcerative colitis (UC)\cite{26}. Top SNP rs3774937 was associated with differential DNA methylation at 413 CpG sites across the genome, 92\% of which showed the same direction of the effect, i.e. lower methylation associated with the minor allele (binomial \(P=2.72\times10^{-69}\)). Of those 380 CpG sites with lower methylation, 147 (38.7\%) overlap with a nuclear factor kappaB (NFkB) transcription factor binding site (2.75-fold enrichment, \(P=5.3\times10^{-32}\)), as derived from ENCODE NFkB ChIP-seq data in blood cell types (Figure 4c). Three motif enrichment analyses (Homer, PWMEnrich, DEEPbind)\textsuperscript{21–23} corroborated the enrichment of NFkB binding motifs for the 413 CpG sites (Figure 4c). Notably, SNP rs3774937 is located in the first intron of \textit{NFKB1} and we found that the minor allele was associated with higher \textit{NFKB1} expression (Figure 4a). Of the 413 \textit{trans}-CpGs, 64 were eQTMs and revealed a coherent gene network (Figure 4d) that was enriched for immunological processes related to \textit{NFKB1} function\cite{27} (Figure 4e). Taken together, these results support the idea that the minor allele of rs3774937, which is associated with increased UC risk, decreases DNA methylation \textit{in trans} by increasing \textit{NFKB1} expression \textit{in cis}.

The same analysis approach indicated that the 779 \textit{trans}-methylation effects of rs8060686 (associated with various phenotypes including metabolic syndrome\textsuperscript{28} and coronary heart disease\textsuperscript{29}) were mediated by altered CTCF binding which mapped 315 kb from the \textit{trans}-meQTL SNP. We observed a strong CTCF ChIP-seq enrichment with 603/779 \textit{trans}-CpGs overlapping with CTCF binding (\(P=1.6\times10^{-232}\)) and enrichment for CTCF motifs (Figure 5). Of these \textit{trans}-CpGs, only 13 were observed previously in lymphocytes\textsuperscript{17}. Hence, the minor allele of rs8060686 increased DNA methylation \textit{in trans} which could be attributed to a lower CTCF gene expression \textit{in cis}.

We found another example of this phenomenon: 228 \textit{trans}-meQTL effects of 4 SNPs on chromosome 10, mapping near \textit{NKX2-3} and implicated in inflammatory bowel disease\textsuperscript{26}, were strongly enriched for \textit{NKX2} transcription factor motifs and associated with \textit{NKX2-3} expression. Again, a negative correlation was observed: the minor allele of rs111900140 decreased DNA methylation \textit{in trans} at \textit{NKX2-3} binding sites and increased \textit{NKX2-3} gene expression \textit{in cis} (Supplementary Fig. 6).
Figure 4. An imbalance in effect direction of trans-meQTLs implies involvement of transcription factors. a, Each dot represents a SNP with at least 10 trans-meQTL effects. The x-axis shows the number of trans-effects where the minor allele decreases methylation, whereas the y-axis shows an increase in methylation. SNPs with a multitude of effects of which many have the same allelic direction often exhibit evidence for a cis-eQTL on a transcription factor (colored dots), and an overrepresentation of trans-CpGs overlapping binding sites for that transcription factor. b, Depiction of the NFKB1 gene and rs3774937, associated with ulcerative colitis and an increased expression of NFKB1 for the risk and minor allele C. Boxes show the median and IQR. Whiskers show the outer quartile plus 1.5 times the IQR. c, In addition to influencing NFKB1 expression, rs3774937 also relates to DNA methylation at 413 CpGs in trans, decreasing methylation levels at 93% of affected CpG sites (dark grey). Many of the CpG sites (37.3%) overlap with NFKB binding sites (3.8-fold enrichment, P-value=5.3x10^-32) (outer chart). d, Gene network of the eQTM genes associated with 72 of the 413 CpGs (17.4%), that are showing a trans-meQTL and an trans-eQTL (in red). NFKB1 is depicted in blue, illustrations of the observed trans-meQTL (left plot) and trans-eQTL effects (right plot) of rs3774937. e, Top pathways as identified by DEPICT for which the genes in d were overrepresented. Many of the identified pathways were inflammation-related, in line with the inflammatory nature of ulcerative colitis.
A height locus harboring 4 SNPs and is associated with 267 trans-CpGs implicated a role for ZBTB38 in mediating trans-meQTL effects (Supplementary Fig. 7). In contrast to the aforementioned TFs that are all transcriptional activators, ZBTB38 is a transcriptional repressor and its expression was positively correlated with methylation in trans, which is in line with our observation that eQTM in repressed regions are enriched for positive correlations. Finally, the trans-methylation effects of rs7216064 (64 trans-CpGs), associated with lung carcinoma, preferentially occurred at regions binding CTCF, while the SNP was located in the BPTF gene, which is known to occupy CTCF binding sites (Supplementary Fig. 8).

The possibility to link trans-meQTL effects to an association of TF expression in cis and concomitant differential methylation in trans at the respective binding site is limited to TFs for which ChIP-seq data or motif information is available. In order to make inferences on TFs for which such data is not yet available, we ascertained whether trans-meQTLs SNPs were more often associated with TF gene expression in cis as compared with SNPs without a trans-meQTL effect. We observed that 13.1% of the GWAS SNPs that produced trans-meQTLs also affect TF gene expression in cis, whereas only 4.5% of the GWAS SNPs without a trans-meQTLs affects TF gene expression in cis (Fisher’s exact $P=6.6 \times 10^{-13}$).
Here we report that one third of known disease- and trait-associated SNPs has downstream methylation effects in trans and often are associated with multiple regions across the genome. Our data suggest that the biological mechanism underlying trans-meQTLs commonly involves a local effect on the expression of a nearby TF that influences DNA methylation at the distal binding sites of that particular TF. The direction of downstream methylation effects is remarkably consistent for each SNP and indicates that decreased DNA methylation is a signature of increased binding of transcriptional activators. As such, our study reveals previously unrecognized functional consequences of disease variants in non-coding regions. These can be looked up online (see URLs), and will provide leads for experimental follow-up.

Methods

Cohort descriptions

The five cohorts used in our study are described briefly below. The number of samples per cohort and references to full cohort descriptions can be found in Supplementary Table 1.

CODAM

The Cohort on Diabetes and Atherosclerosis Maastricht (CODAM) consists of a selection of 547 subjects from a larger population-based cohort. Inclusion of subjects into CODAM was based on a moderately increased risk to develop cardiometabolic diseases, such as type 2 diabetes and/or cardiovascular disease. Subjects were included if they were of Caucasian descent and over 40 years of age and additionally met at least one of the following criteria: increased BMI (>25), a positive family history of type 2 diabetes, a history of gestational diabetes and/or glycosuria, or use of anti-hypertensive medication.

LifeLines-DEEP

The LifeLines-DEEP (LLD) cohort is a sub-cohort of the LifeLines cohort. LifeLines is a multidisciplinary prospective population-based cohort study examining the health and health-related behaviours of 167,729 individuals living in the northern parts of The Netherlands using a unique three-generation design. It employs a broad range of investigative procedures assessing the biomedical, socio-demographic, behavioural, physical and psychological factors contributing to health and disease in the general population. A subset of 1,500 LifeLines participants also take part in LLD. For these participants, additional molecular data is generated, allowing for a more thorough investigation of the association between genetic and phenotypic variation.

LLS

The aim of the Leiden Longevity Study (LLS) is to identify genetic factors influencing longevity and examine their interaction with the environment in order to develop interventions to increase health at older ages. To this end, long-lived siblings of European descent were recruited together with their offspring and their offspring’s partners, on the condition that at least two long-lived siblings were alive at the time of ascertainment. For men the age criteria was 89 or older, for women age 91 or over. These criteria led to the ascertainment of 944 long-lived siblings from 421 families, together with 1,671 of their offspring and 744 partners.

NTR

The Netherlands Twin Register was established in 1987 to study the extent to which genetic and environmental influences cause phenotypic differences between individuals. To this end, data from twins and their families (nearly 200,000 participants) from all over the Netherlands are collected, with a focus on health, lifestyle, personality, brain development, cognition, mental health, and aging.
The Rotterdam Study is a single-centre, prospective population-based cohort study conducted in Rotterdam, the Netherlands. Subjects were included in different phases, with a total of 14,926 men and women aged 45 and over included as of late 2008. The main objective of the Rotterdam Study is to investigate the prevalence and incidence of and risk factors for chronic diseases to contribute to a better prevention and treatment of such diseases in the elderly.

Genotype data

Data generation

Genotype data was generated for each cohort individually. Details on the methods used can be found in the individual papers (CODAM: van Dam et al.; LLD: Tigchelaar et al.; LLS: Deelen et al.; NTR: Willemsen et al.; RS: Hofman et al.).

Imputation and QC

For each cohort separately, the genotype data were harmonized towards the Genome of the Netherlands using Genotype Hamonizer and subsequently imputed per cohort using Impute2 using GoNL reference panel (v5). Quality control was also performed per cohort. We removed SNPs based on imputation info-score (<0.5), HWE ($P<10^{-4}$), call rate (<95%) and minor allele frequency (>0.05), resulting in 5,206,562 SNPs that passed quality control in each of the datasets.

Methylation data

Data generation

For the generation of genome-wide DNA methylation data, 500 ng of genomic DNA was bisulfite modified using the EZ DNA Methylation kit (Zymo Research, Irvine, California, USA) and hybridized on Illumina 450k arrays according to the manufacturer’s protocols. The original IDAT files were generated by the Illumina iScan BeadChip scanner. We collected methylation data for a total of 3,841 samples. Data was generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (see URLs).

Probe remapping and selection

We remapped the 450K probes to the human genome reference (HG19) to correct for inaccurate mappings of probes and identify probes that mapped to multiple locations on the genome. Details on this procedure can be found in Bonder et al. (2014). Next, we removed probes with a known SNP (GoNL, MAF > 0.01) at the single base extension (SBE) site or CpG site. Lastly, we removed all probes on the sex chromosomes, leaving 405,709 high quality methylation probes for the analyses.

Normalization and QC

Methylation data was processed using a custom pipeline based on the pipeline developed by Tost & Toulemat. First, we used methylumi to extract the data from the raw IDAT files. Next, we removed incorrectly mapped probes and checked for outlying samples using the first two principal components (PCs) obtained using principal component analysis (PCA). None of the samples failed our quality control checks, indicating high quality data. Following quality control, we performed background correction and probe type normalization as implemented in DASEN. Normalization was performed per cohort, followed by quantile normalization on the combined data to normalize the differences per cohort. We used mix-up mapper to identify sample mix-ups between genotype and DNA methylation data, detecting and correcting 193 mix-ups. Lastly, in order to correct for known and unknown confounding sources of variation in the methylation data and increase statistical power, we removed the first components which were not affected by genetic information (22 PCs) from the methylation data using methodology we have successfully used in trans-eQTL and meQTL analyses.
RNA sequencing

Total RNA from whole blood was deprived of globin using Ambion’s GLOBIN clear kit and subsequently processed for sequencing using Illumina’s Truseq version 2 library preparation kit. Paired-end sequencing of 2x50bp was performed using Illumina’s HiSeq2000, pooling 10 samples per lane. Finally, read sets per sample were generated using CASAVA, retaining only reads passing Illumina’s Chastity Filter for further processing. Data was generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (see URLs).

Initial QC was performed using FastQC, v0.10.1 (See URLs), removal of adaptors was performed using cutadapt, v1.1, and Sickle, v1.2 (See URLs) was used to trim low quality ends of the reads (min length 25, min quality 20). The sequencing reads were mapped to human genome (HG19) using STAR, v2.3.125. Gene expression quantification was performed by HTseq-count. The gene definitions used for quantification were based on Ensembl version 71, with the extension that regions with overlapping exons were treated as separate genes and reads mapping within these overlapping parts did not count towards expression of the normal genes.

Expression data on the gene level were first normalized using Trimmed Mean of M-values. Then expression values were log2 transformed, gene and sample means were centred to zero. To correct for batch effects, PCA was run on the sample correlation matrix and the first 25 PCs were removed using methodology that we have used before, details are provided in Zhernakova et al.

Cis-meQTL mapping

In order to determine the effect of nearby genetic variation on methylation levels (cis-meQTLs, here defined as the relationship between a CpG and a SNP no further than 250kb apart), we performed cis-meQTL mapping using 3,841 samples for which both genotype data and methylation data were available. To this end, we calculated the Spearman rank correlation per cohort, followed by meta-analysis using a weighted Z-method described previously. To detect all possible independent SNPs regulating methylation at a single CpG-site we regressed out all primary cis-meQTL effects and then performed cis-meQTL mapping for the same CpG-site to find secondary cis-meQTL. We repeated this in a stepwise fashion until no more independent cis-meQTL were found.

To filter out potential false positive cis-meQTLs caused by SNPs affecting the binding of a probe on the array, we filtered the cis-meQTLs effects by removing any CpG-SNP pair for which the SNP was located in the probe. In addition, all other CpG-SNP pairs for which the SNP was outside the probe, but in LD (R² > 0.2 or D’ > 0.2) with a SNP inside the probe were also removed. We tested for LD between SNPs in the probe and in the surrounding cis area in the individual genotype datasets, as well as in GoNL v5, in order to be as strict as possible in marking a QTL as true positive.

To correct for multiple testing, we empirically controlled the false discovery rate (FDR) at 5%. For this, we compared the distribution of observed P-values to the distribution obtained from performing the analysis on permuted data. Permutation was done by shuffling the sample identifiers of one data set, breaking the link between, e.g., the genotype data and the methylation or expression data. We repeated this procedure 10 times to obtain a stable distribution of P-values under the null distribution. The FDR was determined by only selecting the strongest effect per CpG in both the real analysis and in the permutations (i.e. probe level FDR < 5%).

Cis-eQTL mapping

For a set of 2,116 BIOS samples we had also generated RNA-seq data. We used this data to identify cis-eQTLs. Cis-eQTL mapping was performed using the same method as cis-meQTL mapping. Details on these eQTLs will be described in a separate paper.
Expression quantitative trait methylation (eQTM) analysis

To identify associations between methylation levels and expression levels of nearby genes (cis-eQTM), we first corrected our expression and methylation data for batch effects and covariates by regressing out the PCs and regressing out the identified cis-meQTLs and cis-eQTLs, to ensure identified associations between CpG sites and gene expression levels were not due to shared genetic effects. We mapped eQTMs in a window of 250Kb around the TSS of a transcript. Further statistical analysis was identical to the cis-meQTL mapping. For this analysis we were able to use a total of 2,101 samples for which both genetic, methylation and gene expression data was available. To correct for multiple testing we controlled the FDR at 5%, the FDR was determined by only selecting the strongest effect per CpG in both the real analysis and in the permutations.

Trans-meQTL mapping

To identify the effects of distal genetic variation with methylation (trans-meQTLs) we used the same 3,841 samples that we had used for cis-meQTL mapping. To focus our analysis and limit the multiple testing burden, we restricted our analysis to SNPs that have been previously found to be significantly correlated to traits and diseases. We extracted these SNPs from the NHGRI genome-wide association study (GWAS) catalogue, used recent GWAS studies not yet in the NHGRI GWAS catalogue and studies on the Immunochip and Metabochip platform that are not included in the NHGRI GWAS catalogue (Supplemental file 1). We compiled this list of SNPs in December 2014. Per SNP we only investigated CpG sites that mapped at least 5 Mb from the SNP or on other chromosomes. Before mapping trans-meQTLs, we regressed out the identified cis-meQTLs to increase the statistical power of trans-meQTL detection (as done previously for trans-eQTLs3) and to avoid designating an association as trans that may be due to long-range LD (e.g. within the HLA region). To ascertain the stability of the trans-meQTLs we also performed the trans-mapping using uncorrected data cell-type proportions corrected methylation data. In addition, we performed meQTL mapping on SNPs known to influence the cell type proportions in blood18,19.

To filter out potential false positive trans-meQTLs due to cross-hybridization of the probe, we remapped the methylation probes with very relaxed settings, identical to Westra et al.3, with the difference that we only accepted mappings if the last bases of the probe including the SBE site were accurately mapped to the alternative location. If the probe mapped within our minimal trans-window, 5 Mb from the SNP, we removed the effect as being a false positive trans-meQTL.

We controlled the false-discovery rate at 5%, identical to the aforementioned cis-meQTL analysis.

Trans-eQTL mapping

To check if the trans-meQTL effects also showed in gene expression levels, we annotated the CpGs with a trans-meQTL to genes using our eQTMs. Using the 2,101 samples for which both genotype and gene expression data were available, we performed trans-eQTL mapping, associating the SNPs known to be associated with DNA methylation in trans with their corresponding eQTM genes.

Annotations and enrichment tests

Annotation of the CpGs was performed using Ensembl54 (v70), UCSC Genome Browser55 and data from the Epigenomics Roadmap Project.56 We used the Epigenomics Roadmap annotation for the SBE site of the methylation site using 27 blood cell types. We used both the histone mark information and the chromatin marks in blood-related cell types only, as generated by the Epigenomics Roadmap Project. Summarizing the information over the 27 blood cell types was done by counting presence of histone-marks in all the cell types and scaling the abundance, i.e. if the mark is bound in all cell types the score would be 1 if it would be present in none of the blood cell types the score would be 0.
To calculate enrichment of meQTLs or eQTMs for any particular genomic context, we used logistic regression because this allowed us to account for covariates such as CpG methylation variation. For cis-meQTLs, we used the variability of DNA methylation, the number of SNPs tested, and the distance to the nearest SNP per CpG as covariates. For all other analyses we used only the variability in DNA methylation as a covariate.

We used transcription factor ChIP-seq data from the ENCODE-project for blood-related cell lines (narrow peak data). We overlapped CpG locations with ChIP-seq signals and performed a Fisher exact test to determine whether the trans-meQTL probes associated with a SNP were overlapping a ChIP-seq region more often than other trans-meQTL probes.

Enrichment of known sequence motifs among trans-CpGs was assessed by PWMEnrich package in R, Homer and DEEPbind. For PWMEnrich, hundred base pair sequences around the interrogated CpG site were used, and as a background set we used the top CpGs from the 50 permutations used to determine the FDR threshold of the trans-meQTLs. For Homer the default settings for motif enrichment identification were used, and the same CpGs derived from the permutations were used as a background. For DEEPbind we used both the permutation background like described for Homer and the permutations background as described for PWMEnrich.

Using data published by Rao et al. we were able to intersect the trans-meQTLs with information about the 3D structure of the human genome using combined Hi-C data for both inter- and intra-chromosomal data at 1Kb and the quality threshold of E30 in the GM12878 lymphoblastoid cell line. Both the trans-meQTL SNP and trans-meQTL probes were put in the relevant 1Kb block, and for these blocks we looked up the chromosomal contact value in the measurements by Rao et al. Surrounding the trans-meQTLs SNPs, we used a LD window that spans maximally 250Kb from the trans-meQTL SNP and had a minimal R2 of 0.8. If a Hi-C contact between the SNP block and the CpG-site was indicated, we flagged the region as a positive for Hi-C contacts. As a background, we used the combinations found in our 50 permuted trans-meQTL analyses, taking for each permutation the top trans-meQTLs that were similar in size to the real analysis.

eQTM direction prediction

We predicted the direction of the eQTM effects using both a decision tree and a naive Bayes model (as implemented by Rapid-miner v6.3). We built the models on the strongest eQTMs (FDR<9.73x10^-6). For the decision tree we used a standard cross-validation set-up using 20 folds. For the naive Bayes model we used a double loop cross-validation: performance was evaluated in the outer loop using 20-fold cross-validation, while feature selection (using both backward elimination and forward selection) took place in the inner loop using 10-fold cross-validation. Details about the double-loop cross-validation can be found in Ronde et al. During the training of the model, we balanced the two classes making sure we had an equal number of positively correlating and negatively correlating CpG-gene combinations, by randomly sampling a subset of the overrepresented negatively correlating CpG-gene combination group. We chose to do so to circumvent labelling al eQTMs as negative, since this is the class were the majority of the eQTMs are in.

In the models we used CpG-centric annotations: overlap with epigenomics roadmap chromatin states, histone marks and relations between the histone marks, GC content surrounding the CpG-site and relative locations from the CpG-site to the transcript.

DEPICT

To investigate whether there was biological coherence in the trans-meQTLs identified for the NFkB1 locus, we performed gene-set enrichment analysis for the genes near the trans-CpG sites of the UC genetic risk factor (which maps in the NFkB1 locus). To do so, we adapted DEPICT, a pathway enrichment analysis method that we originally have developed for
GWAS. Instead of defining loci with genes by using the top associated SNPs (as is done when analysing GWAS data), we used the eQTM information to empirically link trans-CpGs to genes (that map close to the CpGs). Within the DEPICT gene set enrichment, significance is determined by using a background set of genes. As a background in the adapted DEPICT enrichment analyses we matched our background to the results from the actual trans-meQTL and eQTM analyses: the matching was performed by generating a set of background CpGs (and corresponding correlating eQTM genes), by selecting an equal number of CpGs for which we had found trans-meQTL effects with SNPs that map outside the NFKB locus. By doing so we ensure that the characteristics of these background CpGs are the same as the real NFKB trans-meQTL CpGs, both in terms of CpG variance and the requirement that they also show a significant correlation with expression levels of genes close to the CpG (i.e. a cis-eQTM), ensuring that the corresponding input genes for DEPICT have the same expression variation distribution in the actual NFKB analysis and in the background. Subsequent pathway enrichment analysis was conducted as described before27, and significance was determined by controlling the false discovery rate at 5%.

URLs


Accession codes

All results can be queried using our dedicated QTL browser, see URLs. Raw data was submitted to the European Genome-phenome Archive (EGA), under accession EGAS00001001077.

Author contributions

BTH, PACtH, JBJvM, Al, RJ and LF formed the management team of the BIOS consortium. DIB, RP, JVD, JZH, MMJVG, CDAS, CJHvdK, CGS, CW, LF, AZ, EFG, PES, MB, JD, DvH, JHV, LHvdB, CMvd, BAH, AI, AGU managed and organized the biobanks. JBJvM, PMJ, MV, HEDS, MV, RvdB, JvR and NL generated RNA-seq and Illumina 450k data. HM, MvI, MvG, JB, DVZ, RJ, PvtH, PD, IN, PACtH, BTH and MM were responsible for data management and the computational infrastructure. MJB, RL, MV, DVZ, RS, IJ, Mvl, PD, Fvd, MvG, WA, SMK, MAS, EWvZ, RJ, PACtH, LF and BTH performed the data analysis. MJB, RL, LF and BTH drafted the manuscript. D.V.Z, M.M., P.D. and M.V. contributed equally. A.I., R.J. and J.B.J.M. contributed equally

Competing financial interests

The authors declare no competing financial interests.
References


Acknowledgements

This work was performed within the framework of the Biobank-Based Integrative Omics Studies (BIOS) Consortium funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007). Samples were contributed by LifeLines, the Leiden Longevity Study, the Netherlands Twin Registry (NTR), the Rotterdam studies, the Genetic Research in Isolated Populations program, the Codam study and the PAN study. We thank the participants of all aforementioned biobanks and acknowledge the contributions of the investigators to this study (Supplemental Acknowledgements). This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. L.F. is supported by a grant from the Dutch Research Council (ZonMW-VIDI 917.14.374) and is supported by FP7/2007–2013, grant agreement 259867, and by an ERC Starting Grant, grant agreement 637640 (ImmRisk).
Additional files

The following supplemental data files are available with the online version of this paper.

**Supplementary Figure 1** Density of distances between CpG-site and strongest associated meQTL SNP

**Supplementary Figure 2** Relation between methylation variation and meQTL associated CpGs

**Supplementary Figure 3** Characterization of cis-meQTLs

**Supplementary Figure 4** Characterization of cis-eQTM in relation to the direction of the eQTM effect

**Supplementary Figure 5** Trans-meQTLs identified for a risk factor for inflammatory bowel disease, rs11190140, and the overlap with NKX2-3

**Supplementary Figure 6** Trans-meQTLs identified for a risk factor for height, rs6763931, and the overlap with ZBTB38

**Supplementary Figure 7** Trans-meQTLs identified for a risk factor related to lung carcinoma, rs7216064, and overlap with BPTF

**Supplementary Table 1** Descriptions and number of samples per cohort.

**Supplementary Table 2** Number of independent cis-meQTLs per QTL mapping round

**Supplementary Table 3** GWAS SNPs tested for trans-meQTLs

**Supplementary Table 4** Replication of lymphocytes trans-meQTLs in blood and vice-versa

**Supplementary Table 5** Results of trans-meQTL in non-corrected data

**Supplementary Table 6** Results of trans-meQTL in blood-cell composition corrected data

**Supplementary Table 7** Results of trans-meQTL mapping on Blood cell composition related SNPs

**Supplementary Table 8** Trans-meQTL effects replicated in expression

**Supplementary note 1** Supplementary results & Acknowledgements