From correlation to causation: Data-driven exploration of transcriptional regulation using population genomics

Luijk, R.

3 DISEASE VARIANTS ALTER TRANSCRIPTION FACTOR LEVELS AND METHYLATION LEVELS OF THEIR BINDING SITES

M.J. Bonder*, **René Luijk***, D.V. Zhernakova, M. Moed, P. Deelen, M. Vermaat, M. van Iterson, F. van Dijk, M. van Galen, J. Bot, R.C. Slieker, P.M. Jhamai, M. Verbiest, H.E. Suchiman, M. Verkerk, R. van der Breggen, J. van Rooij, N. Lakenberg, W. Arindrarto, S.M. Kielbasa, I. Jonkers, P. van 't Hof, I. Nooren, M. Beekman, J. Deelen, D. van Heemst, A. Zhernakova, E.F. Tigchelaar, M.A. Swertz, A. Hofman, A.G. Uitterlinden, R. Pool, J. van Dongen, J.J. Hottenga, C.D. Stehouwer, C.J. van der Kallen, C.G. Schalkwijk, L.H. van den Berg, E.W. van Zwet, H. Mei, Y. Li, M. Lemire, T.J. Hudson, BIOS Consortium, P.E. Slagboom, C. Wijmenga, J.H. Veldink, M.M. van Greevenbroek, C.M. van Duijn, D.I. Boomsma, A. Isaacs, R. Jansen, J.B. van Meurs, P.A.C. 't Hoen, L. Franke, B.T. Heijmans

* *Contributed equally*

Nature Genetics, **49**(1):131-138 (2017)

Main

Most disease-associated genetic variants are noncoding, making it challenging to design experiments to understand their functional consequences [Manolio, 2010; Visscher et al., 2012]. Identification of expression quantitative trait loci (eQTLs) has been a powerful approach to infer the downstream effects of disease-associated variants, but most of these variants remain unexplained [Westra et al., 2013; Wright et al., 2014]. The analysis of DNA methylation, a key component of the epigenome [Bernstein et al., 2007; Mill and Heijmans, 2013], offers highly complementary data on the regulatory potential of genomic regions [Gutierrez-Arcelus et al., 2013; Tsankov et al., 2015]. Here we show that disease-associated 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 sets from 3,841 Dutch individuals, we identified 1,907 established trait-associated SNPs that affect the methylation levels of 10,141 different CpG sites *in trans* (false discovery rate (FDR) < 0.05). These included SNPs that affect both the expression of a nearby transcription factor (such as *NFKB1*, *CTCF* and *NKX2-3*) and methylation of its respective binding site across the genome. *Trans* methylation QTLs effectively expose the downstream effects of disease-associated variants.

To systematically study the role of DNA methylation in explaining the downstream effects of genetic variation, we analyzed genome-wide genotype and DNA methylation in whole blood from 3,841 samples from five Dutch biobanks [Tigchelaar et al., 2015; van Greevenbroek et al., 2011; Schoenmaker et al., 2006; Willemsen et al., 2013; Hofman et al., 2013] (Figure 3.1, Supplementary Table 1 and Supplementary Note). We found *cis* methylation quantitative trait locus (meQTL) effects for 34.4% of all 405,709 CpGs tested (n $= 139,566$ at a CpG-level FDR of 5%, $P < 1.38 \times 10^{-4}$), typically with a short physical distance between the SNP and CpG (median distance = 10 kb; Supplementary Figure 3.1). By regressing out the effect of the primary meQTL for each of these CpGs and repeating the *cis*-meQTL mapping, we observed up to 16 independent *cis*-meQTLs for each CpG site (Supplementary Table 2), totaling 272,037 independent *cis*-meQTL effects. We found that few factors determine whether a CpG site shows a *cis*-meQTL effect other than variance in the methylation levels of the CpG site involved (Supplementary Figures 2 and 3). The proportion of variance in methylation explained by SNPs, however, is typically small (Supplementary Figure 3.3b). When accounting for this strong effect of CpG variation, we found only modest enrichments and depletions of *cis*-meQTL CpG sites in CpG island and genic annotations (Supplementary Figure 3e) or when using annotations for biological function based on chromatin segmentations of 27 blood cell types (Figure 3.2a).

We contrasted these modest functional enrichments to those of CpGs whose methylation levels correlated with gene expression in *cis* (that is, expression quantitative trait methylation (eQTM)) by generating RNA-seq data for 2,101 of 3,841 individuals in our study. Using a conservative approach that maximally accounts for potential biases (Online Methods), we identified 12,809 unique

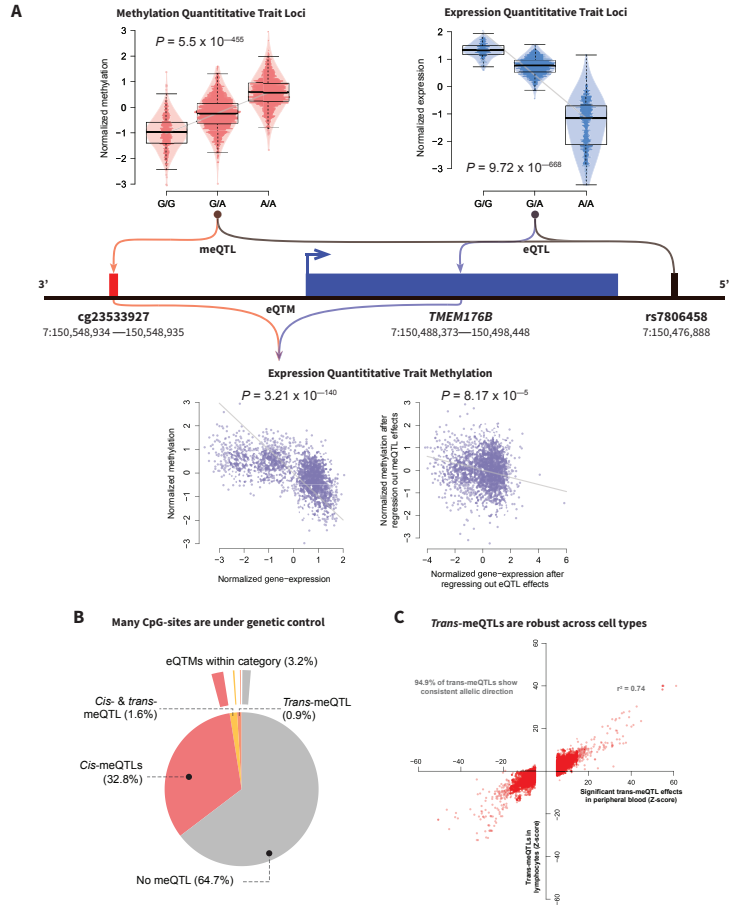

Figure 3.1: (**a**) In the illustration, the relationships between a SNP, DNA methylation at nearby CpGs and associations with the gene itself are shown. Boxes represent the median and interquartile range (IQR); whiskers extend to the outer quartile plus 1.5 times the IQR. The top left plot shows the observed meQTL between cg23533927 and rs7806458. The top right plot shows the observed eQTL between *TMEM176B* and rs7806458. The observed methylation-expression association (eQTM) between *TMEM176B* and cg23533927 is shown below the gene. The bottom left plot 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 in the bottom right plot. (**b**) Two overlaid pie charts. The inner chart indicates the proportion of tested CpGs harboring meQTLs. Over 35% of all tested CpGs show evidence of harboring a meQTL, either *in cis* or *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.

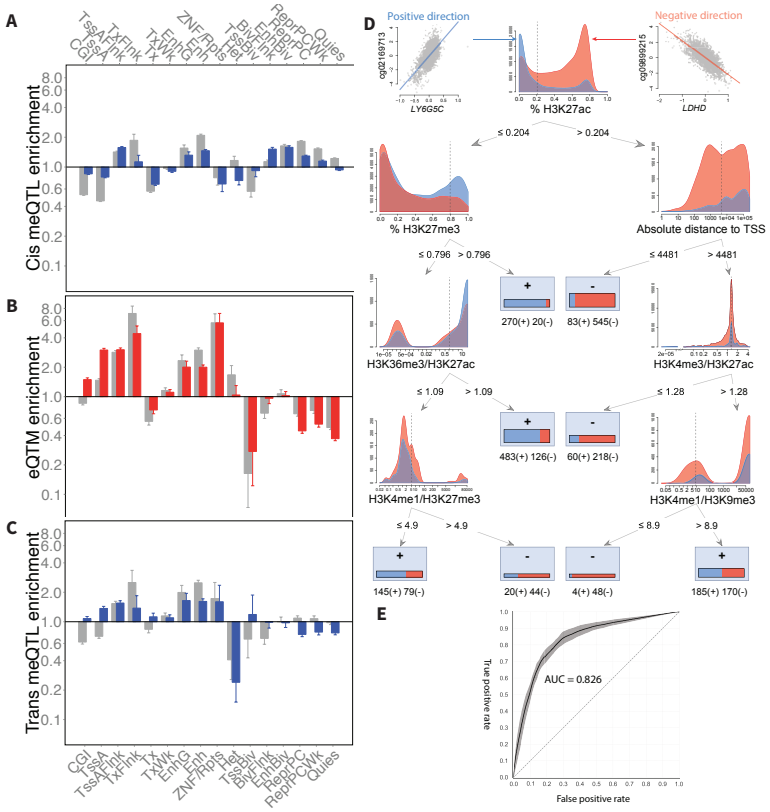

Figure 3.2: (**a-c**) Over- or under-representation of CpGs in predicted chromatin states for *cis*-meQTLs (**a**), eQTMs (**b**) and *trans*-meQTLs (**c**). Gray bars reflect uncorrected enrichment; colored bars reflect enrichment after correction for factors influencing the likelihood of harboring a meQTL or eQTM, including methylation variability. Bar graphs show odds ratios; error bars, 95% confidence intervals. CGI, CpG island; TssA, active TSS; TssAFlnk, flanking active TSS; TxFlnk, transcribed at a gene's 5′ or 3′ end; 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 direction of the effects of eQTMs. Each subplot shows the distributions for positive (blue) and negative (red) associations for that subset of the data. Dashed vertical lines correspond to the optimal split used by the algorithm. The boxes in the leaves show the number of positive and negative effects in each of the leaves. (**e**) Receiver operator characteristic curve showing the performance of the decision tree. AUC, area under the receiver operating curve.

CpGs that correlated with 3,842 unique genes in *cis* (CpG-level FDR < 0.05). eQTMs were enriched for mapping to active regions, for example, 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 3.2b). The majority of eQTMs showed the canonical negative correlation with transcriptional activity (69.2%), but a substantial minority of correlations were positive (30.8%), in line with recent evidence that DNA methylation does not always negatively correlate with gene expression [Hu et al., 2013]. As expected, negatively correlated eQTMs were enriched in active regions such as active TSSs (3.7-fold enrichment, $P = 9.5 \times 10^{-202}$). Positive correlations primarily occurred in repressed regions (for example, Polycomb-repressed regions, 3.4-fold enrichment, $P = 5.8 \times 10^{-103}$) (Supplementary Figure 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 FDR $< 9.7 \times 10^{-6}$, n $= 5,137$), using data on histone marks and distance relative to genes, could predict the direction with an area under the curve of 0.83 (95% confidence interval, $0.78 - 0.87$) (Figure 3.2d,e).

We next ascertained whether *trans*-meQTLs are biologically informative, as previous *trans*-eQTL mapping studies demonstrated that identifying *trans* expression effects provides a powerful tool to uncover and understand the downstream biological effects of disease-associated SNPs [Westra et al., 2013; Yao et al., 2015; Huan et al., 2015]. We focused on 6,111 SNPs that were previously associated with complex traits and diseases ('trait-associated SNPs'; Online Methods and Supplementary Table 3). We observed that one-third of these trait-associated SNPs (1,907 SNPs; 31.2%) affected methylation *in trans* at 10,141 CpG sites, totaling 27,816 SNP-CpG combinations (FDR < 0.05, $P < 2.6 \times 10^{-7}$; Figure 3.3a). This represents a fivefold increase in the number of CpG sites affected as compared with a previous *trans*-meQTL mapping study [Lemire et al., 2015]. We evaluated whether the trait-associated SNPs themselves were likely to underlie the *trans* effects or whether the associations could be attributed to other SNPs in moderate linkage disequilibrium (LD). Of the 1,907 trait-associated 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 are indeed the driving force behind many of the *trans*-meQTLs. Of note, because of the sparse coverage of the Illumina HumanMethylation450 BeadChip, 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 [Lemire et al., 2015]. Of the 18,764 overlapping *trans*-meQTLs, 94.9% had a consistent allelic direction in the replication data (Figure 3.1e and Supplementary Table 4). This indicates that the identified *trans*-meQTLs are robust and are not caused by differences in cell type composition. Further analysis of SNPs known to influence blood cell composition [Orru et al., 2013; Roederer et al., 2015] showed no or only few effects in *trans* and alternative adjustments of the methylation data corroborated the stability of the *trans* effects, with both approaches indicating a limited influence of cell type composition (Supplementary Tables 5, 6, 7 and Supplementary Note).

After identifying *trans*-meQTLs, we assessed whether their respective SNPs also affected the expression of the genes associated with the CpGs *in trans*. By overlaying the *trans*-meQTLs and *cis*-eQTMs, we could link 436 SNPs to 850 genes, totaling 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 effect given the directions of effect for the *trans*-meQTL and *cis*-eQTM (Supplementary Table 8).

In contrast to *cis*-meQTL CpGs, *trans*-meQTL CpGs showed substantial functional enrichment: they were enriched around TSSs and depleted in heterochromatin (Figure 3.2c) and were strongly enriched for being an eQTM (1,913 CpGs (18.9%), 5.2-fold enrichment, $P = 2.3 \times 10^{-101}$). Among the 1,907 trait-associated SNPs that made up the *trans*-meQTLs, there was an over-representation of GWAS-identified SNPs associated with immune- and cancer-related traits (Figure 3.3a). The large majority of *trans*-meQTLs were interchromosomal (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 22 autosomal chromosomes (*trans* bands; Figure 3.3b).

We subsequently studied the nature of these *trans*-meQTLs. Using high-resolution Hi-C data [Rao et al., 2014], we identified 720 SNP-CpG pairs (including 402 CpG sites and 172 SNPs) among the *trans*-meQTLs that overlapped with an interchromosomal contact, which is 2.9-fold more than expected by chance ($P = 3.7 \times 10^{-126}$; Figure 3.3a,b). The enrichment for Hi-C interchromosomal contacts remained after removing SNPs that were responsible for *trans* bands ($P = 1.7 \times 10^{-61}$). Hence, interchromosomal contacts may produce associations between SNPs and CpGs *in trans*. To characterize the 720 SNP-CpG pairs overlapping with interchromosomal contacts, we examined motif enrichment using three motif enrichment analysis tools (HOMER, PWMEnrich and DEEPbind; Heinz et al. [2010]; Alipanahi et al. [2015]). These analyses showed that the 402 CpG sites involved frequently overlapped with binding sites for CTCF, RAD21 and SMC3 ($P = 2.3 \times 10^{-5}$, $P = 3.5 \times 10^{-5}$ and $P = 5.1 \times 10^{-5}$, respectively), factors known to regulate chromatin architecture [Zuin et al., 2014; Splinter et al., 2006]. 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 *trans*-meQTLs reflected the effect of differential transcription factor binding for transcription factors that mapped close to the SNPs. The rationale for this hypothesis is that binding of transcription factors has been linked to changes in local DNA methylation, primarily loss of methylation upon transcription factor binding and gain of methylation after loss of transcription factor occupancy [Gutierrez-Arcelus et al., 2013; Tsankov et al., 2015]. This model suggests that *trans*-meQTLs may be attributed to SNPs affecting the expression of a transcription factor *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 was associated with multiple CpG sites in *trans* (at least 10, $n = 305$), the direction of the association of the SNP was consistently skewed toward either increased or decreased DNA methylation. On average, 76% of the CpGs for each *trans*-meQTL SNP displayed the same direction of effect (50%

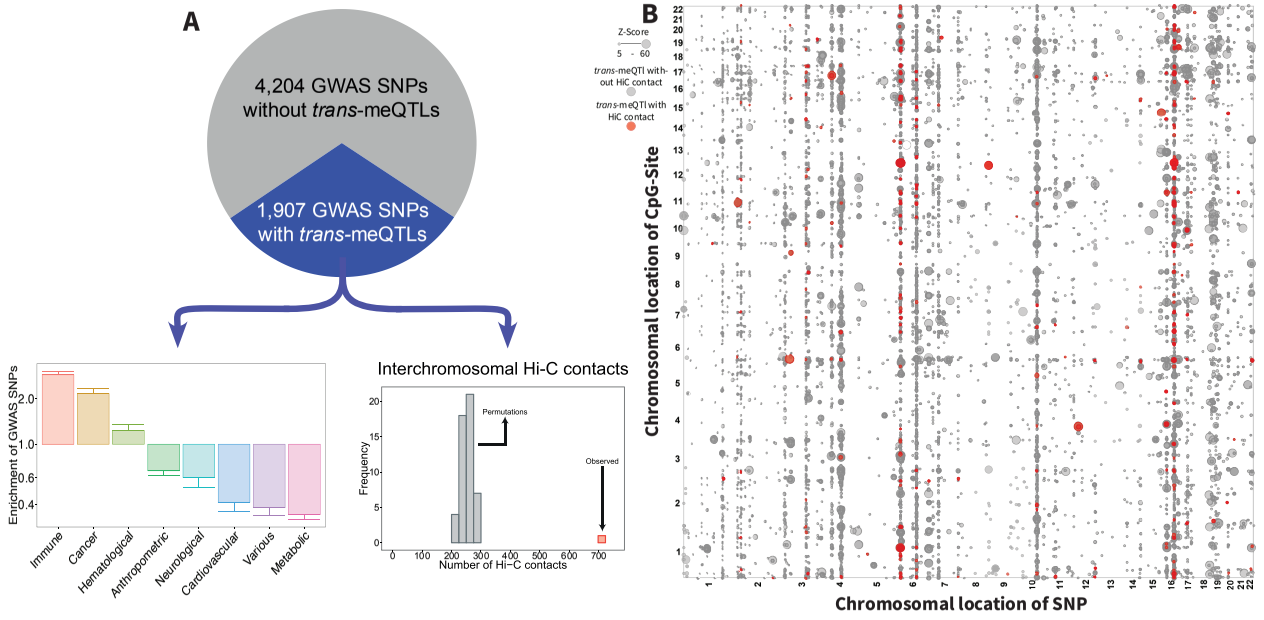

Figure 3.3: (**a**) Distribution of the tested trait-associated SNPs influencing DNA methylation *in trans*. Over 1,900 (31.2%) of all tested SNPs have downstream effects on DNA methylation. Bottom left, for the associated GWAS SNPs, we show the over-representation of SNPs with *trans*-meQTLs in different GWAS trait categories, where the y-axis shows the odds ratio and the bars depict the error margin. Bottom right, Hi-C contacts are over-represented among *trans*-meQTLs. Gray bars show the number of Hi-C contacts using permutated data, and the red bar corresponds to the actually observed number in our data. (**b**) Dot plot depicting the *trans*-meQTLs. Effect strength is reflected by the size of each dot. Red dots correspond to *trans*-meQTLs that overlap with a Hi-C contact site. Several SNPs with widespread *trans*-meQTLs show interchromosomal contacts across the genome, further implicating an important role for those SNPs in development of the associated trait.

expected, $P = 10^{-111}$; Figure 3.4a). A significant skew in the direction of the allelic effect was present for 59.7% of the 305 individual SNPs with at least 10 *trans*-meQTL effects, and this proportion increased to 95.2% for the 104 SNPs with at least 50 *trans*-meQTL effects (binomial $P < 0.05$), suggesting that differential transcription factor binding might explain a substantial fraction of *trans*-meQTLs.

To explore this mechanism further, we combined ChIP-seq data on transcription factor binding at CpGs with the expression effects *in cis* of SNPs to directly examine the involvement of transcription factors in mediating *trans*-meQTLs. Among the trait-associated SNPs influencing at least 10 CpGs *in trans* (n = 305), we identified 13 *trans*-meQTL SNPs with strong support for a role of transcription factors (Figure 3.4a).

The most striking example was a locus on chromosome 4 (Figure 3.4b), where two SNPs (rs3774937 and rs3774959; in strong LD) were associated with ulcerative colitis [Jostins et al., 2012]. The top SNP, rs3774937, was associated with differential DNA methylation at 413 CpG sites across the genome, 92% of which showed the same direction of effect—that is, lower methylation—associated with the minor allele (binomial $P = 2.72 \times 10^{-69}$). Of the 380 CpG sites with lower methylation, 147 (38.7%) overlapped with a nuclear factor (NF)-κB transcription factor binding site (2.75-fold enrichment, $P = 5.3 \times 10^{-32}$), as derived from Encyclopedia of DNA Elements (ENCODE) *NF-κB* ChIP-seq data in blood cell types (Figure 3.4c). Three motif enrichment analysis tools (HOMER, PWMEnrich and DEEPbind) [Heinz et al., 2010; Alipanahi et al., 2015] corroborated the enrichment of *NF-κB*-binding motifs for the 413 CpG sites (Figure 3.4c). Notably, SNP rs3774937 is located in the first intron of *NFKB1*, and we found that the minor allele was associated with higher *NFKB1* expression (Figure 3.4a). Of the 413 CpGs *in trans*, 64 were eQTMs and showed a coherent gene network (Figure 3.4d) that was enriched for immunological processes related to *NFKB1* function [Pers et al., 2015] (Figure 3.4e). Taken together, these results support the idea that the minor allele of rs3774937, which is associated with increased risk of ulcerative colitis, decreases DNA methylation *in trans* by increasing *NFKB1* expression *in cis*.

The same analysis approach indicated that the 779 methylation effects of rs8060686 *in trans* (associated with various phenotypes, including metabolic syndrome [Kristiansson et al., 2012] and coronary heart disease [Lettre et al., 2011]) were mediated by altered CTCF binding, which mapped 315 kb from the *trans*-meQTL SNP. We observed strong CTCF ChIP-seq enrichment (603 of the 779 CpGs *in trans* overlapping with CTCF binding; $P = 1.6 \times 10^{-232}$) and enrichment for CTCF motifs (Figure 3.5). Of these *trans* CpGs, only 13 were observed previously in lymphocytes [Lemire et al., 2015]. Hence, the minor allele of rs8060686 increased DNA methylation *in trans*, which could be attributed to lower *CTCF* gene expression *in cis*.

We found another example of this phenomenon: 228 *trans*-meQTL effects of four SNPs on chromosome 10, mapping near *NKX2-3* and implicated in inflammatory bowel disease [Jostins et al., 2012], were strongly enriched for NKX2 transcription factor motifs and associated with *NKX2-3* expression. Again, a negative correlation was observed, in which the minor allele of rs11190140 decreased DNA methylation *in trans* at NKX2-3-binding sites and increased *NKX2-*

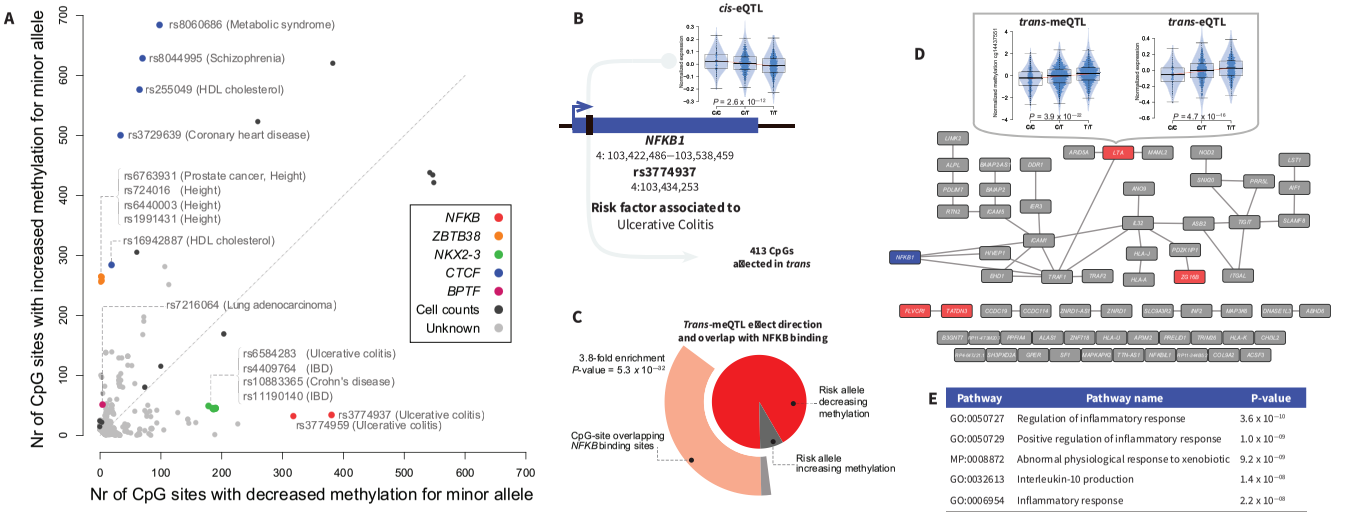

Figure 3.4: (**a**) Each dot represents a SNP with at least ten *trans*-meQTL effects. The x axis shows the number of *trans* effects where the minor allele decreases methylation, and the y axis shows the number of *trans* effects where the minor allele increases methylation. SNPs with a multitude of effects of which many have the same allelic direction often exhibit evidence of a *cis*-eQTL on a transcription factor (colored dots) and an over-representation of *trans*-CpGs overlapping binding sites for that transcription factor. (**b**) Depiction of the *NFKB1* gene and rs3774937, for which the risk and minor allele C is associated with ulcerative colitis and increased expression of *NFKB1*. Boxes show the median and IQR; whiskers extend to 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 the affected CpG sites (dark gray). Outer chart, many of the CpG sites (37.3%) overlap with NF-κB-binding sites (3.8-fold enrichment, $P = 5.3 \times 10^{-32}$). (**d**) Gene network of the eQTM genes associated with 72 of the 413 CpGs (17.4%) that show a *trans*-meQTL and a *trans*-eQTL (in red). *NFKB1* is depicted in blue. The illustrations above show theobserved *trans*-meQTL (left plot) and *trans*-eQTL (right plot) effects of rs3774937. (**e**) Top pathways as identified by DEPICT for which the genes in **d** were over-represented. Many of the identified pathways are related to inflammation, in line with the inflammatory nature of ulcerative colitis.

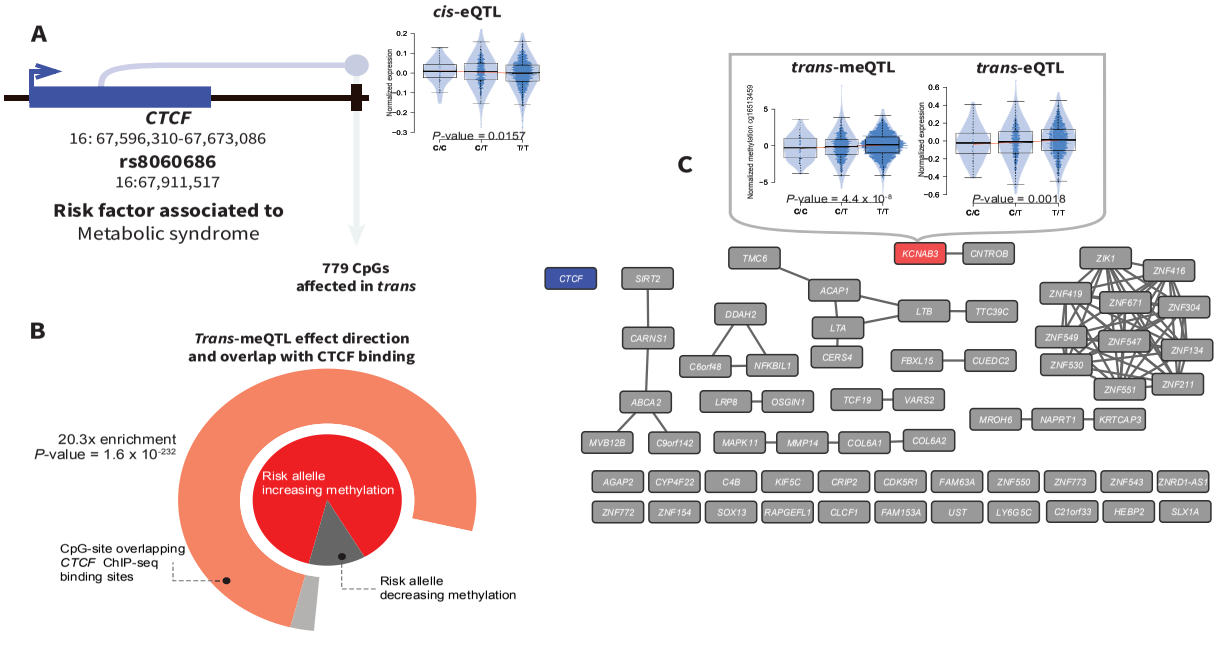

Figure 3.5: (**a**) Depiction of the *CTCF* gene and rs8060686, associated with metabolic syndrome. The plot shows increased expression of *CTCF* for the risk allele C. (**b**) In addition to influencing *CTCF* expression, rs8060686 also influences DNA methylation at 779 CpGs *in trans*, increasing methylation levels at 87.7% of the affected CpG sites (dark gray). Outer chart, many of the CpG sites (77.4%) overlap with CTCF-binding sites (20.3-fold enrichment, $P = 1.6x10^{-232}$). (**c**) Gene network of the genes associated with 60 of the 779 CpGs (7.7%) with a *trans*-meQTL. In the top part of the figure, there is an illustration of overlapping *trans*-meQTL (left) and *trans*-eQTL effects (right) for rs8060686.

3 gene expression *in cis* (Supplementary Figure 5).

A height-associated locus [Soranzo et al., 2009] harboring four SNPs and associated with 267 *trans* CpGs implicated a role for *ZBTB38* in mediating *trans*-meQTL effects (Supplementary Figure 6). In contrast to the aforementioned transcription factors, which are all transcriptional activators, *ZBTB38* is a transcriptional repressor [Filion et al., 2006; Sasai and Defossez, 2009] and its expression was positively correlated with methylation *in trans*, in line with our observation that eQTMs in repressed regions are enriched for positive correlations. Finally, the methylation effects *in trans* of rs7216064 (64 *trans* CpGs), associated with lung carcinoma [Shiraishi et al., 2012], preferentially occurred at regions binding CTCF, while the SNP was located in the *BPTF* gene, which encodes a protein known to occupy CTCF-binding sites [Qiu et al., 2015] (Supplementary Figure 7).

The possibility of linking *trans*-meQTL effects to an association with transcription factor expression *in cis* and concomitant differential methylation *in trans* at the respective binding site for the transcription factor is limited to transcription factors for which ChIP-seq data or motif information is available. To make inferences on transcription factors for which such data are not yet available, we ascertained whether *trans*-meQTL SNPs were more often associated with transcription factor gene expression *in cis* as compared with SNPs without a *trans*-meQTL effect. We observed that 13.1% of the trait-associated SNPs that produced *trans*-meQTLs also affected transcription factor gene expression *in cis*, whereas only 4.5% of the trait-associated SNPs without a *trans*-meQTL affected transcription factor 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 have downstream effects on methylation *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 transcription factor that influences DNA methylation at the distal binding sites of that particular transcription factor. 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 identifies the previously unrecognized functional consequences of disease-associated variants in noncoding regions. These can be viewed 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) [van Greevenbroek et al., 2011] consists of a selection of 547 subjects from a larger population-based cohort [van Dam et al., 2001]. Inclusion of subjects into CODAM was based on a moderately increased risk of developing cardiometabolic diseases, such as type 2 diabetes and/or cardiovascular disease. Subjects were included if they were of European ancestry and over 40 years of age and additionally met at least one of the following criteria: increased body mass index (BMI; > 25), a positive family history for type 2 diabetes, a history of gestational diabetes and/or glycosuria, or use of antihypertensive medication.

LifeLines-DEEP

The LifeLines-DEEP (LLD) cohort [Tigchelaar et al., 2015] is a subcohort of the LifeLines cohort [Scholtens et al., 2015]. LifeLines is a multidisciplinary prospective population-based cohort study examining the health and health-related behaviors 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 biomedical, sociodemographic, behavioral, 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 [Tigchelaar et al., 2015]. For these participants, additional molecular data are generated, allowing for a more thorough investigation of the association between genetic and phenotypic variation.

LLS

The aim of the Leiden Longevity Study (LLS) [Schoenmaker et al., 2006] is to identify genetic factors influencing longevity and examine their interaction with the environment as a means 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 criterion was 89 years or older; for women, the age criterion was 91 years or older. 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 (NTR) [Willemsen et al., 2013; Boomsma et al., 2002, 2008] 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.

RS

The Rotterdam Study [Hofman et al., 2013] is a single-center, 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 years 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 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., 2001]; LLD [Tigchelaar et al., 2015]; LLS [Deelen et al., 2014a]; NTR [Willemsen et al., 2013]; RS [Hofman et al., 2013]).

Imputation and QC

For each cohort separately, the genotype data were harmonized toward the Genome of the Netherlands (GoNL) using Genotype Hamonizer [Deelen et al., 2014b] and subsequently imputed per cohort using Impute2 [Howie et al., 2009] using GoNL [Deelen et al., 2014c] 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 data sets.

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) 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 Touleimat and Tost [2012]. 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 [Pidsley et al., 2013]. Normalization was performed per cohort, followed by quantile normalization on the combined data to normalize the differences per cohort. We used mix-up mapper [Westra et al., 2011] 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 [Westra et al., 2013; Fehrmann et al., 2011] and meQTL analyses [Touleimat and Tost, 2012].

RNA sequencing

Total RNA from whole blood was depleted of globin transcripts using the Ambion GLOBIN clear kit and subsequently processed for sequencing using the Illumina TruSeq version 2 library preparation kit. Paired-end sequencing of 2×50-bp reads was performed using the Illumina HiSeq 2000 platform, pooling ten samples per lane. Finally, read sets were generated for each sample using CASAVA, retaining only reads passing the Illumina Chastity Filter for further processing. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (see URLs).

Initial quality control was performed using FastQC v0.10.1 (see URLs), removal of adaptors was performed using cutadapt [Martin, 2011] (v1.1) and Sickle v1.2 (see URLs) was used to trim low-quality ends from the reads (min length 25, min quality 20). Sequencing reads were mapped to the human genome (hg19) using STAR [Dobin et al., 2013] 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 toward expression of the normal genes.

Expression data on the gene level were first normalized using trimmed mean of M values [Robinson and Oshlack, 2010]. Then, expression values were log2 transformed, and gene and sample means were centered to zero. To correct for batch effects, principal-component analysis (PCA) was run on the sample

correlation matrix and the first 25 principal components were removed using methodology that we have used before [Westra et al., 2013; Fehrmann et al., 2011]; details are provided in Zhernakova et al. [2016].

Cis-meQTL mapping

To determine the effect of nearby genetic variation on methylation levels (*cis*-meQTL, here defined as the relationship between a CpG and a SNP no further than 250 kb 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 for each cohort, followed by meta-analysis using a weighted Z-method described previously [Westra et al., 2013]. 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*-meQTLs. We repeated this in a stepwise fashion until no more independent *cis*-meQTLs 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*-meQTL effects by removing any CpG-SNP pairs 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 ($r^2 > 0.2$ or D$\prime > 0.2$) with a SNP inside the probe were also removed. We tested for LD between SNPs in probes and in surrounding *cis* areas in the individual genotype data sets, as well as in GoNL v5, to be as strict as possible in marking a QTL as a true positive.

To correct for multiple testing, we empirically controlled the 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 performed by shuffling the sample identifiers of one data set, thereby breaking the link between, for example, the genotype data and the methylation or expression data. We repeated this procedure ten times to obtain a stable distribution of P-values under the null distribution. The FDR was determined by only selecting the strongest effect for each CpG [Westra et al., 2013] in both the real analysis and the permutations (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 are described in a separate paper [Zhernakova et al., 2016].

Expression quantitative trait methylation analysis

To identify associations between methylation levels and the expression levels of nearby genes (*cis*-eQTMs), we first corrected our expression and methylation data for batch effects and covariates by regressing out the principal components

and regressing out the identified *cis*-meQTLs and *cis*-eQTLs, to ensure that the associations identified between CpG sites and gene expression levels were not due to shared genetic effects. We mapped the eQTMs in a window of 250 kb around the TSS of a transcript. Further statistical analysis was identical to that for *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 were available. To correct for multiple testing, we controlled the FDR at 5%; the FDR was determined by only selecting the strongest effect for each CpG [Westra et al., 2013] in both the real analysis and the permutations.

Trans-meQTL mapping

To identify the effects of distal genetic variation on 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 previously been found to be significantly correlated with traits and diseases. We extracted these SNPs from the NHGRI GWAS catalog and also used recent GWAS not yet in the NHGRI GWAS catalog and studies on the Immunochip and Metabochip platforms that are not included in the NHGRI GWAS catalog. We compiled this list of SNPs in December 2014. For each 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*-eQTLs [Westra et al., 2013]) and to avoid designating an association as *trans* that might be due to long-range LD (for example, within the human leukocyte antigen (HLA) region). To ascertain the stability of the *trans*-meQTLs, we also performed *trans* mapping using uncorrected methylation data and data corrected for cell type proportions. In addition, we performed meQTL mapping on SNPs known to influence cell type proportions in blood [Orru et al., 2013; Roederer et al., 2015].

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 those used in Westra et al. [2013], 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 FDR at 5%, identical to in the aforementioned *cis*-meQTL analysis.

Trans-eQTL mapping

To check whether *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 SNPs known to be associated with DNA methylation in *trans* with their corresponding eQTM genes.

Annotation and enrichment tests

Annotation of CpG sites was performed using Ensembl [Flicek et al., 2013] (v70), the UCSC Genome Browser [Kent et al., 2002] and data from the Epigenomics Roadmap project [Kundaje et al., 2015]. We used 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 carried out by counting the presence of histone marks in all the cell types and scaling the abundance: that is, the score would be 1 if a mark is bound in all cell types, whereas the score would be 0 if it is present in none of the blood cell types.

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 in DNA methylation, the number of SNPs tested and the distance to the nearest SNP for each 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's exact test to determine whether the *trans*-meQTL probes associated with a SNP overlapped a ChIP-seq region more often than other *trans*-meQTL probes.

Enrichment of known sequence motifs among *trans*-CpGs was assessed using the PWMEnrich package in R, HOMER [Heinz et al., 2013] and DEEPbind [Alipanahi et al., 2015]. For PWMEnrich, the 100-bp sequence around each interrogated CpG site was 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 the identification of motif enrichment were used, and the same CpG sites derived from the permutations were used as background. For DEEPbind, we used both the permutation background as described for HOMER and the permutation background as described for PWMEnrich.

Using data published by Rao et al. [2014], 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 intrachromosomal data at 1 kb and the quality threshold of E30 in the GM12878 LCL. Both the *trans*-meQTL SNPs and *trans*-meQTL probes were put in the relevant 1-kb blocks, and for these blocks we looked up the chromosomal contact value in the measurements by Rao *et al*. Surrounding the *trans*-meQTL SNPs, we used an LD window that spanned maximally 250 kb from the *trans*-meQTL SNP and had a minimal r^2 value of 0.8. If a Hi-C contact was indicated between a SNP block and a CpG site, we flagged the region as positive for Hi-C contacts. As background, we used the combinations

found in our 50 permutated *trans*-meQTL analyses, taking for each permutation the top *trans*-meQTLs that were similar in size to those from the real analysis.

Prediction of eQTM direction

We predicted the direction of eQTM effects using both a decision tree and a naive Bayes model (as implemented by Rapid-miner v6.3 [Hofmann and Klinkenberg, 2013]). We built the models on the strongest eQTMs (FDR $< 9.73 \times 10^{-6}$). For the decision tree, we used a standard cross-validation setup with 20 folds. For the naive Bayesian model, we used 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 tenfold cross-validation. Details about double-loop cross-validation can be found in de Ronde et al. [2014]. 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 over-represented negatively correlating CpG-gene combination group. We chose to do so to circumvent labeling al eQTMs as negative, as this is the class to which the majority of the eQTMs belonged.

In the models, we used CpG-centric annotations: overlap with Epigenomics Roadmap chromatin states, histone marks and relationships 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 ulcerative colitis genetic risk factor (which maps in the *NFKB1* locus). To do so, we adapted DEPICT [Pers et al., 2015], a pathway enrichment analysis method that we originally developed for GWAS. Instead of defining loci with genes by using the top associated SNPs (as is done when analyzing GWAS data), we used the eQTM information to empirically link *trans*-CpGs to genes (that map close to the CpGs). Within DEPICT gene set enrichment, significance is determined by using a background set of genes. As background in the adapted DEPICT enrichment analyses, we matched our background to the results from the actual *trans*-meQTL and eQTM analyses: 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 *NFKB1* locus. By doing so, we ensured that the characteristics of these background CpGs were the same as those for the real *NFKB1 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 (that is, a *cis*-eQTM), ensuring that the corresponding input genes for DEPICT had the same expression variation distribution in the actual *NFKB1* analysis and in the background. Subsequent

pathway enrichment analysis was conducted as described before [Pers et al., 2015], and significance was determined by controlling the FDR at 5%.

URLs

All results can be queried using our dedicated QTL browser at http://www.genenetwork.nl/biosqtlbrowser. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (http://www.glimDNA.org/). Cohort webpages are as follows: LifeLines, http://lifelines.nl/lifelines-research/general; Leiden Longevity Study, http://www.healthy-ageing.nl/ and http://www.leidenlangleven.nl/; Netherlands Twin Registry, http://www.tweelingenregister.org/; Rotterdam Studies, http://www.erasmusmc.nl/epi/research/The-Rotterdam-Study/; Genetic Research in Isolated Populations program, http://www.epib.nl/research/geneticepi/research.html#gip; CODAM study, http://www.carimmaastricht.nl/; PAN study, http://www.alsonderzoek.nl/. Software used included the following: FastQC, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; Sickle, https://github.com/najoshi/sickle; PWMEnrich: PWM enrichment analysis v.4.6.0, https://bioconductor.riken.jp/packages/3.2/bioc/html/PWMEnrich.html.

Accession codes

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

References

Alipanahi, B. et al. [2015]. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning, *Nat. Biotechnol.* **33**: 831–838.

Bernstein, B. E., Meissner, A. and Lander, E. S. [2007]. The mammalian epigenome, *Cell* **128**: 669–681.

Bonder, M. J. et al. [2014]. Genetic and epigenetic regulation of gene expression in fetal and adult human livers, *BMC Genomics* **15**: 860.

Boomsma, D. I. et al. [2002]. Netherlands twin register: a focus on longitudinal research, *Twin Res.* **5**: 401–406.

Boomsma, D. I. et al. [2008]. Genome-wide association of major depression: description of samples for the gain major depressive disorder study: Ntr and nesda biobank projects, *Eur. J. Hum. Genet.* **16**: 335–342.

de Ronde, J. J. et al. [2014]. Breast cancer subtype specific classifiers of response to neoadjuvant chemotherapy do not outperform classifiers trained on all subtypes, *PLoS One* **9**: e88551.

Deelen, J. et al. [2014a]. Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age, *Hum. Mol. Genet.* **23**: 4420–4432.

Deelen, P. et al. [2014b]. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration, *BMC Res. Notes* **7**: 901.

Deelen, P. et al. [2014c]. Improved imputation quality of low-frequency and rare variants in european samples using the 'genome of the netherlands', *Eur. J. Hum. Genet.* **22**: 1321–1326.

Dobin, A. et al. [2013]. Star: ultrafast universal rna-seq aligner, *Bioinformatics* **29**: 15–21.

Fehrmann, R. S. N. et al. [2011]. Trans-eqtls reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the hla, *PLoS Genet.* **7**: e1002197.

Filion, G. J. P. et al. [2006]. A family of human zinc finger proteins that bind methylated dna and repress transcription, *Mol. Cell. Biol.* **26**: 169–181.

Flicek, P. et al. [2013]. Ensembl 2013, *Nucleic Acids Res.* **41**: D48–D55.

Gutierrez-Arcelus, M. et al. [2013]. Passive and active dna methylation and the interplay with genetic variation in gene regulation, *eLife* **2**.

Heinz, S. et al. [2010]. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities, *Mol. Cell* **38**: 576–589.

Heinz, S. et al. [2013]. Effect of natural genetic variation on enhancer selection and function, *Nature* **503**: 487–492.

Hofman, A. et al. [2013]. The rotterdam study: 2014 objectives and design update, *Eur. J. Epidemiol.* **28**: 889–926.

Hofmann, M. and Klinkenberg, R. [2013]. *Rapid Miner Data Mining Use Cases and Business Analytics Applications* .

Howie, B. N. et al. [2009]. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLoS Genet* **5**(6): e1000529.

Hu, S. et al. [2013]. Dna methylation presents distinct binding sites for human transcription factors, *eLife* **2**: e00726.

Huan, T. et al. [2015]. A meta-analysis of gene expression signatures of blood pressure and hypertension, *PLoS Genet.* **11**: e1005035.

Jostins, L. et al. [2012]. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease, *Nature* **491**: 119–124.

Kent, W. J. et al. [2002]. The human genome browser at ucsc, *Genome Res.* **12**: 996–1006.

Kristiansson, K. et al. [2012]. Genome-wide screen for metabolic syndrome susceptibility loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits, *Circ Cardiovasc Genet* **5**: 242–249.

Kundaje, A. et al. [2015]. Integrative analysis of 111 reference human epigenomes, *Nature* **518**: 317–330.

Lemire, M. et al. [2015]. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci, *Nat. Commun.* **6**: 6326.

Lettre, G. et al. [2011]. Genome-wide association study of coronary heart disease and its risk factors in 8,090 african americans: the nhlbi care project, *PLoS Genet.* **7**: e1001300.

Manolio, T. A. [2010]. Genomewide association studies and assessment of the risk of disease, *N. Engl. J. Med.* **363**: 166–176.

Martin, M. [2011]. Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal* **17**: 10–12.

Mill, J. and Heijmans, B. T. [2013]. From promises to practical strategies in epigenetic epidemiology, *Nat Rev Genet* **14**(8): 585–594.

Orru, V. et al. [2013]. Genetic variants regulating immune cell levels in health and disease, *Cell* **155**: 242–256.

Pers, T. H. et al. [2015]. Biological interpretation of genome-wide association studies using predicted gene functions, *Nat. Commun.* **6**: 5890.

Pidsley, R. et al. [2013]. A data-driven approach to preprocessing illumina 450k methylation array data, *BMC Genomics* **14**: 293.

Qiu, Z. et al. [2015]. Functional interactions between nurf and ctcf regulate gene expression, *Mol. Cell. Biol.* **35**: 224–237.

Rao, S. S. P. et al. [2014]. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping, *Cell* **159**: 1665–1680.

Robinson, M. D. and Oshlack, A. [2010]. A scaling normalization method for differential expression analysis of rna-seq data, *Genome Biol.* **11**: R25.

Roederer, M. et al. [2015]. The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis, *Cell* **161**: 387–403.

Sasai, N. and Defossez, P. A. [2009]. Many paths to one goal?: The proteins that recognize methylated dna in eukaryotes, *Int. J. Dev. Biol.* **53**: 323–334.

Schoenmaker, M. et al. [2006]. Evidence of genetic enrichment for exceptional survival using a family approach: the leiden longevity study, *Eur. J. Hum. Genet.* **14**: 79–84.

Scholtens, S. et al. [2015]. Cohort profile: Lifelines, a three-generation cohort study and biobank, *Int. J. Epidemiol.* **44**: 1172–1180.

Shiraishi, K. et al. [2012]. A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the japanese population, *Nat. Genet.* **44**: 900–903.

Soranzo, N. et al. [2009]. Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size, *PLoS Genet.* **5**: e1000445.

Splinter, E. et al. [2006]. Ctcf mediates long-range chromatin looping and local histone modification in the [beta]-globin locus, *Genes Dev.* **20**: 2349–2354.

Tigchelaar, E. F. et al. [2015]. Cohort profile: Lifelines deep, a prospective, general population cohort study in the northern netherlands: study design and baseline characteristics, *BMJ Open* **5**: e006772.

Touleimat, N. and Tost, J. [2012]. Complete pipeline for infinium([reg]) human methylation 450k beadchip data processing using subset quantile normalization for accurate dna methylation estimation, *Epigenomics* **4**: 325–341.

Tsankov, A. M. et al. [2015]. Transcription factor binding dynamics during human es cell differentiation, *Nature* **518**: 344–349.

van Dam, R. M. et al. [2001]. Parental history of diabetes modifies the association between abdominal adiposity and hyperglycemia, *Diabetes Care* **24**: 1454–1459.

van Greevenbroek, M. M. J. et al. [2011]. The cross-sectional association between insulin resistance and circulating complement c3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the codam study), *Eur. J. Clin. Invest.* **41**: 372–379.

Visscher, P. M. et al. [2012]. Five years of gwas discovery, *Am. J. Hum. Genet.* **90**: 7–24.

Westra, H. et al. [2011]. Mixupmapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects, *Bioinformatics* **27**(15): 2104–2111.

Westra, H. et al. [2013]. Systematic identification of trans eqtls as putative drivers of known

disease associations, *Nat Genet* **45**(10): 1238–1243.

Willemsen, G. et al. [2013]. The adult netherlands twin register: twenty-five years of survey and biological data collection, *Twin Res. Hum. Genet.* **16**: 271–281.

Wright, F. A. et al. [2014]. Heritability and genomics of gene expression in peripheral blood, *Nat. Genet.* **46**: 430–437.

Yao, C. et al. [2015]. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes, *Circulation* **131**: 536–549.

Zhernakova, D. V. et al. [2016]. Identification of context-dependent expression quantitative trait loci in whole blood, *Nat. Genet.* .

Zuin, J. et al. [2014]. Cohesin and ctcf differentially affect chromatin architecture and gene expression in human cells, *Proc. Natl. Acad. Sci. USA* **111**: 996–1001.