

# 1 Disease variants alter transcription factor 2 levels and methylation of their binding sites

3  
4 Marc Jan Bonder<sup>1,\*</sup>, René Luijk<sup>2,\*</sup>, Daria V. Zhernakova<sup>1,\*\*\*</sup>, Matthijs Moed<sup>2,\*\*\*</sup>, Patrick  
5 Deelen<sup>1,3,\*\*\*</sup>, Martijn Vermaat<sup>4,\*\*\*</sup>, Maarten van Iterson<sup>2</sup>, Freerk van Dijk<sup>1,3</sup>, Michiel van Galen<sup>3</sup>,  
6 Jan Bot<sup>5</sup>, Roderick C. Slieker<sup>2</sup>, P. Mila Jhamai<sup>6</sup>, Michael Verbiest<sup>3</sup>, H. Eka D. Suchiman<sup>2</sup>, Marijn  
7 Verkerk<sup>6</sup>, Ruud van der Breggen<sup>2</sup>, Jeroen van Rooij<sup>6</sup>, Nico Lakenberg<sup>2</sup>, Wibowo Arindrarto<sup>8</sup>,  
8 Szymon M. Kielbasa<sup>7</sup>, Iris Jonkers<sup>2</sup>, Peter van 't Hof<sup>7</sup>, Irene Nooren<sup>5</sup>, Marian Beekman<sup>2</sup>, Joris  
9 Deelen<sup>2</sup>, Diana van Heemst<sup>9</sup>, Alexandra Zhernakova<sup>1</sup>, Ettje F. Tigchelaar<sup>1</sup>, Morris A. Swertz<sup>1,3</sup>,  
10 Albert Hofman<sup>10</sup>, André G. Uitterlinden<sup>6</sup>, René Pool<sup>11</sup>, Jenny van Dongen<sup>11</sup>, Jouke J. Hottenga<sup>11</sup>,  
11 Coen D.A. Stehouwer<sup>12</sup>, Carla J.H. van der Kallen<sup>12</sup>, Casper G. Schalkwijk<sup>12</sup>, Leonard H. van den  
12 Berg<sup>13</sup>, Erik. W van Zwet<sup>8</sup>, Hailiang Mei<sup>7</sup>, Yang Li<sup>1</sup>, Mathieu Lemire<sup>14</sup>, Thomas J. Hudson<sup>14,15,16</sup>,  
13 the BIOS Consortium, P. Eline Slagboom<sup>2</sup>, Cisca Wijmenga<sup>1</sup>, Jan H. Veldink<sup>13</sup>, Marleen M.J. van  
14 Greevenbroek<sup>12</sup>, Cornelia M. van Duijn<sup>17</sup>, Dorret I. Boomsma<sup>11</sup>, Aaron Isaacs<sup>17,##</sup>, Rick  
15 Jansen<sup>18,##</sup>, Joyce B.J. van Meurs<sup>6,##</sup>, Peter A.C. 't Hoen<sup>4,#</sup>, Lude Franke<sup>1,#</sup>, Bastiaan T. Heijmans<sup>2,#</sup>

16 \* Shared first, \*\* Shared second, ## Shared second last, # Shared last

17 Corresponding authors:

18 Lude Franke ([lude@ludesign.nl](mailto:lude@ludesign.nl)) and Bastiaan T. Heijmans ([bas.heijmans@lumc.nl](mailto:bas.heijmans@lumc.nl))

19 <sup>1</sup> Department of Genetics, University of Groningen, University Medical Centre Groningen,  
20 Groningen, The Netherlands

21 <sup>2</sup> Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden  
22 University Medical Center, Leiden, The Netherlands

23 <sup>3</sup> Genomics Coordination Center, University Medical Center Groningen, University of Groningen,  
24 Groningen, the Netherlands

25 <sup>4</sup> Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

26 <sup>5</sup> SURFsara, Amsterdam, the Netherlands

27 <sup>6</sup> Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands

28 <sup>7</sup> Sequence Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands

29 <sup>8</sup> Medical Statistics Section, Department of Medical Statistics and Bioinformatics, Leiden  
30 University Medical Center, Leiden, The Netherlands

31 <sup>9</sup> Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The  
32 Netherlands

33 <sup>10</sup> Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands

34 <sup>11</sup> Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus  
35 Amsterdam, Amsterdam, The Netherlands

36 <sup>12</sup> Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht  
37 University Medical Center, Maastricht, The Netherlands

38 <sup>13</sup> Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht,  
39 Utrecht, The Netherlands

40 <sup>14</sup>Ontario Institute for Cancer Research, Toronto, Ontario, Canada M5G 0A3

41 <sup>15</sup>Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada M5S 1A1

42 <sup>16</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A1

43 <sup>17</sup> Genetic Epidemiology Unit, Department of Epidemiology, ErasmusMC, Rotterdam, The  
44 Netherlands

45 <sup>18</sup> Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam,  
46 Amsterdam, The Netherlands

47

48 **Most disease associated genetic risk factors are non-coding, making it challenging to design**  
49 **experiments to understand their functional consequences<sup>1,2</sup>. Identification of expression**  
50 **quantitative trait loci (eQTLs) has been a powerful approach to infer downstream effects of**  
51 **disease variants but the large majority remains unexplained.<sup>3,4</sup>. The analysis of DNA**  
52 **methylation, a key component of the epigenome<sup>5</sup>, offers highly complementary data on the**  
53 **regulatory potential of genomic regions<sup>6,7</sup>. However, a large-scale, combined analysis of**  
54 **methylome and transcriptome data to infer downstream effects of disease variants is lacking.**  
55 **Here, we show that disease variants have wide-spread effects on DNA methylation *in trans***  
56 **that likely reflect the downstream effects on binding sites of *cis*-regulated transcription**  
57 **factors. Using data on 3,841 Dutch samples, we detected 272,037 independent *cis*-meQTLs**  
58 **(FDR < 0.05) and identified 1,907 trait-associated SNPs that affect methylation levels of**  
59 **10,141 different CpG sites *in trans* (FDR < 0.05), an eight-fold increase in the number of**  
60 **downstream effects that was known from *trans*-eQTL studies<sup>3,8,9</sup>. *Trans*-meQTL CpG sites**  
61 **are enriched for active regulatory regions, being correlated with gene expression and overlap**  
62 **with Hi-C determined interchromosomal contacts<sup>10,11</sup>. We detected many *trans*-meQTL**  
63 **SNPs that affect expression levels of nearby transcription factors (including *NFKB1*, *CTCF***  
64 **and *NKX2-3*), while the corresponding *trans*-meQTL CpG sites frequently coincide with its**  
65 **respective binding site. *Trans*-meQTL mapping therefore provides a strategy for identifying**  
66 **and better understanding downstream functional effects of many disease-associated**  
67 **variants.**

68 To systematically study the role of DNA methylation in explaining downstream effects of genetic  
69 variation, we analysed genome-wide genotype and DNA methylation in whole blood from 3,841  
70 samples from five Dutch biobanks<sup>12-16</sup> (Figure 1 and Extended Data Table 1). We found *cis*-

71 meQTL effects for 34.4% of all 405,709 tested CpGs (n=139,566 at a CpG-level FDR of 5%,  $P \leq$   
72  $1.38 \times 10^{-4}$ ), typically with a short physical distance between the SNP and CpG (median distance  
73 10 kb, Extended Data Fig. 1). By regressing out primary meQTLs effect for each of these CpGs  
74 and repeating the *cis*-meQTL mapping, we observed up to 16 independent *cis*-meQTLs for these  
75 CpGs (Extended Data Table 2). In total, we identified 272,037 independent *cis*-meQTL effects.  
76 Few factors determine whether a CpG site shows a *cis*-meQTL effect except the variance in  
77 methylation level of the CpG site involved: for the top 10% most variable CpGs, 57.2% showed a  
78 *cis*-meQTL effect, dropping to only 8.1% for the 10% least-variable CpGs (Extended Data Fig. 2,  
79 Extended Data Fig. 3a). The proportion of methylation variance explained by SNPs, however, is  
80 typically small (Extended Data Fig. 3b). When accounting for this strong effect of CpG variation,  
81 we find only modest enrichments and depletions for *cis*-meQTL CpG sites when using CpG island  
82 (CGI) and genic annotation (Extended Data Fig. 3e) or when using annotations of biological  
83 function based on chromatin segmentations of 27 blood cell types (Figure 2a).

84 We contrasted these modest functional enrichments to CpGs whose methylation levels correlates  
85 with gene expression *in cis* (i.e. mapping expression quantitative trait methylations (eQTM), by  
86 generating RNA-seq data for 2,101 out of 3,841 individuals in our study. Using a conservative  
87 approach that maximally accounts for potential biases (i.e. *cis*-meQTL effects, *cis*-eQTL effects,  
88 batch effects and cell heterogeneity effects), we identified 12,809 unique CpGs that correlated to  
89 3,842 unique genes *in cis* (CpG-level FDR  $< 0.05$ ). eQTM were enriched for mapping in active  
90 regions, e.g. in and around active TSSs (3-fold enrichment,  $P = 1.8 \times 10^{-91}$ ) and enhancers (2-fold  
91 enrichment,  $P = 1.1 \times 10^{-139}$ , Figure 2b). Of note, the majority of eQTM showed the canonical  
92 negative correlation with transcriptional activity (69.2%) but a substantial minority of correlations  
93 was positive (30.8%) in line with recent evidence that DNA methylation does not always

94 negatively correlate with gene expression<sup>17</sup>. As expected, negatively correlated eQTM<sup>s</sup> were  
95 enriched in active regions like active TSSs (3.7- fold enrichment,  $P = 9.5 \times 10^{-202}$ ). Positive  
96 correlations primarily occurred in repressed regions (e.g. Polycomb repressed, 3.4-fold  
97 enrichment,  $P = 5.8 \times 10^{-103}$ ) (Extended Data Fig. 4). The sharp contrast between positively and  
98 negatively associated eQTM<sup>s</sup>, enabled us to build a model to predict the direction of the  
99 correlation. A decision tree trained on the strongest eQTM<sup>s</sup> (those with an FDR  $< 9.7 \times 10^{-6}$ ,  
100 n=5,137) using data on histone marks and distance relative to gene, could predict the direction  
101 with an area under the curve of 0.83 (95% confidence interval, 0.78-0.87) (Figure 2d, e).

102 We next ascertained whether *trans*-meQTLs are biologically informative, since previous *trans*-  
103 eQTL mapping studies demonstrated that identifying *trans*-expression effects provide a powerful  
104 tool to uncover and understand downstream biological effects of disease-SNPs<sup>3,8,9</sup>. We focussed  
105 on 6,111 SNPs that were previously associated with complex traits and diseases ('trait-associated  
106 SNPs', see Methods and Extended Data Table 3). We observed that one-third of these trait-  
107 associated SNPs (1,907 SNPs, 31.2%) affect methylation *in trans* at 10,141 CpG sites, totalling  
108 27,816 SNP-CpG combinations (FDR  $< 0.05$ ,  $P < 2.6 \times 10^{-7}$ , Figure 3a). This represents a 5-fold  
109 increase in the number of CpG sites affected as compared with a previous trans-meQTL mapping  
110 study<sup>18</sup>. We evaluated whether the GWAS SNP themselves were likely underlying the *trans*-  
111 effects or that the associations could be attributed to another SNP in moderate LD. Of the 1,907  
112 GWAS SNPs with trans-effects, 1,538 (87.2%) were in strong LD with the top SNP ( $R^2 > 0.8$ ),  
113 indicating that the GWAS SNPs indeed are the driving force behind many of the *trans*-meQTLs.  
114 Of note, due to the sparse coverage of the Illumina 450k array, the true number of CpGs in the  
115 genome that are altered by these trait associated SNPs will be substantially higher. After the  
116 identification of the trans-meQTLs, we assessed if the *trans*-meQTLs also are present in

117 expression. Out of the 2,889 testable trans-eQTLs we identified 8.4% of these effects, 91% of the  
118 cases the effect direction was consistent (Extended Data Table 4).

119 To ascertain stability of our *trans*-meQTLs, we performed a replication analysis in a the set of 1,748  
120 lymphocyte samples<sup>18</sup>: of the 18,764 overlapping *trans*-meQTLs between the datasets that could  
121 be tested, 94.9% had a consistent allelic direction (Figure 1E). 12,098 *trans*-meQTLs were  
122 nominally significant (unadjusted  $P < 0.05$ ), of which 99.87% had a consistent allelic direction.  
123 This indicates that the identified *trans*-meQTLs are robust and not caused by differences in cell-  
124 type composition. (Extended Data Table 5). To further ascertain the stability of the *trans*-meQTLs,  
125 we tested SNPs known to influence blood composition<sup>19,20</sup> for effects on methylation *in trans*,  
126 finding these SNPs show no or only few *trans*-meQTLs whereas widespread *trans*-meQTL effects  
127 were to be expected if our analysis had not properly controlled for blood cell composition  
128 (Extended Data Table 6). Furthermore we linked our GWAS SNPs to the SNPs known to influence  
129 cell proportions and found that only 0.6% of the GWAS SNPs are in high LD with SNPs known  
130 to influence cell proportions. Lastly, we performed *trans*-meQTL mapping on uncorrected and cell  
131 type corrected data see supplemental results and Extended Data Table 7,8.

132 In contrast to *cis*-meQTL CpGs, *trans*-meQTLs CpGs show many functional enrichments: they  
133 are enriched around TSSs and depleted in heterochromatin (Figure 2c) and are strongly enriched  
134 for being an eQTM (1,913 CpGs (18.9%), 5.2-fold,  $P = 2.3 \times 10^{-101}$ ). The 1,907 trait-associated  
135 SNPs that make up the *trans*-meQTLs were overrepresented for immune- and cancer-related traits  
136 (Figure 3b). The large majority of *trans*-meQTLs were inter-chromosomal (93%, 9,429 CpG-SNP  
137 pairs) and included 12 *trans*-meQTLs SNPs (yielding 3,616 unique CpG-SNP pairs) that each  
138 showed downstream *trans*-meQTL effects across all of the 22 autosomal chromosomes (i.e. *trans*-  
139 bands, Figure 3d).

140 We subsequently studied the nature of these *trans*-meQTLs. Using high-resolution Hi-C data<sup>10</sup>,  
141 we identified 720 SNP-CpG pairs (including 402 CpG sites and 172 SNPs) among the *trans*-  
142 meQTLs that overlapped with an inter-chromosomal contact, which is 2.9-fold more frequent than  
143 expected by chance ( $P = 3.7 \times 10^{-126}$ , Figure 3c, d). These Hi-C inter chromosomal enrichments  
144 were not confounded due to SNPs that gave *trans*-meQTLs on many CpG sites (i.e. *trans*-bands):  
145 when removing those *trans*-meQTLs from the analysis, Hi-C enrichments remained highly  
146 significant ( $P = 1.7 \times 10^{-61}$ ). This indicates that some relationships between SNPs and CpGs *in trans*  
147 are explained by inter-chromosomal contacts. In order to characterize the 720 SNP-CpG pairs  
148 overlapping with inter-chromosomal contacts, we performed motif enrichments using three motif  
149 enrichment analyses (Homer, PWMEnrich, DEEPbind)<sup>21-23</sup>. These analyses identified that the 402  
150 CpG sites frequently overlapped with CTCF, RAD21 and SMC3 binding sites ( $P = 2.3 \times 10^{-5}$ ,  $P =$   
151  $3.5 \times 10^{-5}$  and  $P = 5.1 \times 10^{-5}$ , respectively), factors known to affect chromatin architecture<sup>24,25</sup>. This  
152 finding was confirmed by incorporating ChIP-Seq data on CTCF binding (1.8-fold enrichment,  $P$   
153 =  $5.2 \times 10^{-7}$ ).

154 We next tested whether the *trans*-meQTLs reflected the effect of differential transcription factor  
155 (TF) binding of TFs that map close to the SNPs since TF binding has been implicated in  
156 demethylation and loss of TF occupancy with remethylation<sup>6,7</sup>. This suggests that if a SNP allele  
157 increases TF el<sup>s</sup> *in cis*, that *trans*-meQTL effects are likely detectable, and that the SNP allele  
158 likely decreases methylation of these CpG sites. Indeed, we observed that if a SNP affects multiple  
159 CpGs sites *in trans* (at least 10, n=305) that the assessed allele often consistently increased or  
160 decreased methylation *in trans*, in the same direction for, on average, 76% of CpGs per *trans*-  
161 meQTL SNP (expected 50%,  $P=10^{-111}$ ; Figure 4a). This skew in allelic effect direction was present  
162 for 59.7% of the 305 SNPs with at least 10 *trans*-meQTL effects increasing to 95.2% for 104 SNPs

163 with at least 50 *trans*-meQTL effects (binomial test  $P < 0.05$ ), suggesting that differential TF  
164 binding may explain a substantial fraction of *trans*-meQTLs.

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

169 The most striking example was a locus on chromosome 4 (Figure 4b), where two SNPs (rs3774937  
170 and rs3774959, in strong LD) were associated with ulcerative colitis (UC)<sup>26</sup>. Top SNP rs3774937  
171 was associated with differential DNA methylation at 413 CpG sites across the genome, 92% of  
172 which showed the same direction of the effect, i.e. lower methylation associated with the risk allele  
173 (binomial  $P=2.72 \times 10^{-69}$ ). Of those 380 CpG sites with lower methylation, 147 (38.7%) overlap  
174 with a nuclear factor kappaB (NFKB) transcription factor binding site (2.75-fold enrichment,  $P =$   
175  $5.3 \times 10^{-32}$ ), as based on ENCODE NFKB ChIP-seq data in blood cell types (Figure 4c). Three motif  
176 enrichment analyses (Homer, PWMEnrich, DEEPbind)<sup>21-23</sup> also revealed an enrichment of NFKB  
177 binding motifs for the 413 CpG sites thus corroborating the ChIP-seq results. Notably, SNP  
178 rs3774937 is located in the first intron of *NFKB1* and we found that the risk allele was associated  
179 with higher *NFKB1* expression (Figure 4a). Of the 413 *trans*-CpGs, 64 were eQTM and revealed  
180 a coherent gene network (Figure 4d) that was enriched for immunological processes related to  
181 *NFKB1* function<sup>27</sup> (Figure 4e). Taken together, these results support the idea that the rs3774937  
182 UC risk allele decreases DNA methylation *in trans* by increasing *NFKB1* expression *in cis*.

183 The same analysis approach indicated that the *trans*-methylation effects of rs8060686 (linked to  
184 various phenotypes including metabolic syndrome<sup>28</sup> and coronary heart disease<sup>29</sup>, and affecting  
185 779 *trans*-CpGs) were due to CTCF which mapped 315 kb from rs8060686. We observed a strong

186 CTCF ChIP-seq enrichment with 603/779 trans-CpGs overlapping with CTCF binding ( $P = 1.6 \times 10^{-232}$ ) and enriched CTCF motifs (Figure 4a and Extended Data Fig. 5). Of these *trans*-CpGs, only 13  
187 have been observed previously in lymphocytes<sup>18</sup>. We observed that the risk allele increased DNA  
188 methylation *in trans* by decreasing *CTCF* gene expression *in cis*.

190 We found another example of this phenomenon: 228 *trans*-meQTL effects of 4 SNPs on  
191 chromosome 10, mapping near *NKX2-3* and implicated in inflammatory bowel disease<sup>26</sup>, were  
192 strongly enriched for NKX2 transcription factor motifs and associated with *NKX2-3* expression.  
193 The risk alleles decreased DNA methylation *in trans* at *NKX2-3* binding sites by increasing *NKX2-3*  
194 gene expression *in cis* (Extended Data Fig. 6).

195 One height locus<sup>30</sup> contained 4 SNPs which influence 267 *trans*-CpGs and implicate *ZBTB38*  
196 (Extended Data Fig. 7). In contrast to the aforementioned TFs that are transcriptional activators,  
197 *ZBTB38* is a transcriptional repressor<sup>31,32</sup> and its expression was positively correlated with  
198 methylation *in trans*, in line with our observation that eQTM<sub>s</sub> in repressed regions are enriched  
199 for positive correlations. Finally, the *trans*-methylation effects of rs7216064 (64 *trans*-CpGs),  
200 associated with lung carcinoma<sup>33</sup>, preferentially occurred at regions binding CTCF, while the SNP  
201 was located in the *BPTF* gene, known to occupy CTCF binding sites<sup>34</sup> (Extended Data Fig. 8).

202 The possibility to link *trans*-meQTL effects to an association of TF expression *in cis* and  
203 concomitant differential methylation *in trans* at the respective binding site is limited to TFs for  
204 which ChIP-seq data or motif information is available. In order to make inferences on TFs for  
205 which such data is not yet available, we ascertained whether *trans*-meQTLs SNPs were more  
206 often affecting TF gene expression *in cis* as compared with SNPs that were not giving *trans*-  
207 meQTLs. We observed that 13.1% of the GWAS SNPs that gave *trans*-meQTLs also affect TF

208 gene expression in *cis*, whereas only 4.5% of the GWAS SNPs that do not give *trans*-meQTLs  
209 affect TF gene expression in *cis* (Fisher's exact P = 6.6 x 10-13).

210 Here we report that one third of known disease- and trait-associated SNPs has downstream  
211 methylation effects *in trans*, often affecting multiple regions across the genome. The biological  
212 mechanism underlying *trans*-meQTLs often involves a local effect on the transcriptional activity  
213 of nearby TFs that affects DNA methylation at distal binding sites of the corresponding TFs. The  
214 direction of downstream methylation effects is remarkably consistent for each SNP and indicates  
215 that decreased DNA methylation is a signature of increased binding of transcriptional activators.  
216 Our study reveals previously unrecognized functional consequences of disease variants in non-  
217 coding regions. These can be looked up online (<http://www.genenetwork.nl/biosqlbrowser/>), and  
218 provide leads for experimental follow-up.

219 **Figures**

220 Figure 1. Overview of a genomic region around *TMEM176B*, where the relations between a SNP,  
221 DNA methylation at nearby CpGs, and the associations with the gene itself are shown. **a**,  
222 Illustration of a methylation Quantitative Trait Locus (meQTL) **b**, Illustration of an expression  
223 Quantitative Trait Locus (eQTL). **c**, Illustration of methylation-expression association (eQTM).  
224 The figures show how correction for meQTLs may increase detection of such associations. The  
225 left plot shows the data before correction for *cis*-meQTLs, the corrected data in the right figure  
226 shows the meQTL-corrected methylation data. **d**, Two overlaid pie charts. The inner chart  
227 indicates the proportion of tested CpGs harboring meQTLs. Over 35% of all tested CpGs show  
228 evidence for harboring a meQTL, either *in cis* or *in trans*. The outer chart indicates what CpGs are

229 associated with gene expression *in cis* (in total 3.2%). **e**, Replication of peripheral blood *trans-*  
230 *meQTLs* in lymphocytes.

231  
232 Figure 2. **a-c**, Over- or underrepresentation of CpGs for different predicted chromatin states for  
233 *cis-meQTLs*, *trans-meQTLs* and eQTMs. Grey bars reflect uncorrected enrichments, colored bars  
234 reflect enrichments after correction for factors influencing the likelihood of harboring a *meQTL*  
235 or eQTM, including methylation variability. Bar graphs show odds ratios and error bars (95%  
236 confidence interval). CGI: CpG island; TssA: Active TSS; TssAFlnk: Flanking active TSS;  
237 TxFlnk, Transcribed at gene 5' and 3'; Tx: Strong transcription; TxWk: Weak transcription; EnhG:  
238 Genic enhancer; Enh: Enhancer; ZNF/Rpts: ZNF genes and repeats; Het: Heterochromatin;  
239 TssBiv: Bivalent/Poised TSS; BivFlnk: Flanking bivalent TSS/Enhancer; EnhBiv: Bivalent  
240 enhancer. **d**, Decision tree for predicting the effect direction of eQTM. Each subplot shows the  
241 distributions for positive (blue) and negative (red) associations for that subset of the data. Dashed  
242 vertical lines indicate the optimal split used by the algorithm. The boxes in the leaves indicate the  
243 number of positive and negative effects in each of the leaves. **e**, Receiver operator characteristic  
244 curve showing the performance of the decision tree. Figure 3. **a**, Distribution of tested trait-  
245 associated SNPs influencing DNA methylation *in trans*. Over 1,900 SNPs (31.2%) of all tested  
246 SNPs have downstream effects on DNA methylation. **b**, Overrepresentation of SNPs with *trans-*  
247 *meQTLs* in different GWAS trait categories, where the y-axis shows the odds ratio. **c**, Hi-C  
248 contacts are overrepresented among *trans-meQTLs*. Grey bars show the number of Hi-C contacts  
249 using permuted data, while the red bar reflects the actually observed number in our data. **d**, Dot-  
250 plot depicting the *trans-meQTLs*. The effect strength is reflected by the size of the dot. Red dots  
251 indicate an overlap with a Hi-C contact. Several SNPs with widespread *trans-meQTLs* show inter-

252 chromosomal contacts genome-wide, further implicating an important role for those SNPs in the  
253 development of the associated trait.

254 Figure 3. **a**, Distribution of tested trait-associated SNPs influencing DNA methylation *in trans*.  
255 Over 1,900 SNPs (31.2%) of all tested SNPs have downstream effects on DNA methylation. **b**,  
256 Overrepresentation of SNPs with *trans*-meQTLs in different GWAS trait categories, where the y-  
257 axis shows the odds ratio. **c**, Hi-C contacts are overrepresented among *trans*-meQTLs. Grey bars  
258 show the number of Hi-C contacts using permuted data, while the red bar reflects the actually  
259 observed number in our data. **d**, Dot-plot depicting the *trans*-meQTLs. The effect strength is  
260 reflected by the size of the dot. Red dots indicate an overlap with a Hi-C contact. Several SNPs  
261 with widespread *trans*-meQTLs show inter-chromosomal contacts genome-wide, further  
262 implicating an important role for those SNPs in the development of the associated trait.

263

264 Figure 4. **a**, An imbalance in effect direction of *trans*-meQTLs implies involvement of  
265 transcription factors. Each dot represents a SNP with at least 10 *trans*-meQTL effects. The x-axis  
266 shows the number of *trans*-effects where the minor allele increases methylation, whereas the y-  
267 axis shows a decrease in methylation. SNPs with a multitude of effects of which many have the  
268 same allelic direction often exhibit evidence for a *cis*-eQTL on a transcription factor (colored dots),  
269 and an overrepresentation of CpGs *in trans* overlapping with binding sites for that transcription  
270 factor. **b**, Depiction of the *NFKB1* gene and rs3774937, associated with ulcerative colitis. The plot  
271 shows an increased expression of *NFKB1* for the risk allele C. **c**, In addition to influencing *NFKB1*  
272 expression, rs3774937 also influences DNA methylation at 413 CpGs *in trans*, decreasing  
273 methylation levels at 93% of affected CpG sites (dark grey). In addition, many of the CpG sites  
274 (37.3%) overlap with NFKB binding sites (3.8-fold enrichment, *P*-value =  $5.3 \times 10^{-32}$ ), shown in  
275 the outer chart. **d**, Illustrations of meQTL (left plot) and eQTL effects (right plot) of rs3774937 *in*  
276 *trans*. Only SNP-gene combinations were tested where the gene was associated with one of the  
277 413 CpGs with a *trans*-meQTL. **e**, Gene network of the eQTM genes associated with 72 of the 413  
278 CpGs (17.4%), that are showing a *trans*-meQTL (red). NFKB is depicted in blue. Genes also  
279 showing evidence for a *trans*-eQTL effect are shown in red. **f**, Top pathways as identified by  
280 enrichment method DEPICT for which many of the genes in **e** were overrepresented. Many of the  
281 identified pathways were inflammation-related, in line with the inflammatory nature of ulcerative  
282 colitis.

283 **Methods**

284 **Cohort descriptions**

285 The five cohorts used in our study are described briefly below. The number of samples per  
286 cohort and references to full cohort descriptions can be found in Extended data table 1.

287 ***CODAM***

288 The Cohort on Diabetes and Atherosclerosis Maastricht<sup>13</sup> (CODAM) consists of a selection of  
289 547 subjects from a larger population-based cohort.<sup>35</sup> Inclusion of subjects into CODAM was  
290 based on a moderately increased risk to develop cardiometabolic diseases, such as type 2  
291 diabetes and/or cardiovascular disease. Subjects were included if they were of Caucasian descent  
292 and over 40 yrs of age and additionally met at least one of the following criteria: increased BMI  
293 (>25), a positive family history of type 2 diabetes, a history of gestational diabetes and/or  
294 glycosuria, or use of anti-hypertensive medication.

295 ***LifeLines-DEEP***

296 The LifeLines-DEEP (LLD) cohort<sup>12</sup> is a sub-cohort of the LifeLines cohort.<sup>36</sup> LifeLines is a  
297 multi-disciplinary prospective population-based cohort study examining the health and health-  
298 related behaviours of 167,729 individuals living in the northern parts of The Netherlands using a  
299 unique three-generation design. It employs a broad range of investigative procedures assessing  
300 the biomedical, socio-demographic, behavioural, physical and psychological factors contributing  
301 to health and disease in the general population, with a special focus on multi-morbidity and  
302 complex genetics. A subset of 1,500 LifeLines participants also take part in LLD<sup>12</sup>. For these  
303 participants, additional molecular data is generated, allowing for a more thorough investigation  
304 of the association between genetic and phenotypic variation.

305 ***LLS***

306 The aim of the Leiden Longevity Study<sup>14</sup> (LLS) is to identify genetic factors influencing  
307 longevity and examine their interaction with the environment in order to develop interventions to  
308 increase health at older ages. To this end, long-lived siblings of European descent were recruited  
309 together with their offspring and their offspring's partners, on the condition that at least two

310 long-lived siblings were alive at the time of ascertainment. For men the age criteria was 89 or  
311 older, for women age 91 or over. These criteria led to the ascertainment of 944 long-lived  
312 siblings from 421 families, together with 1,671 of their offspring and 744 partners.

313 **NTR**

314 The Netherlands Twin Register<sup>15,37,38</sup> (NTR) was established in 1987 to study the extent to which  
315 genetic and environmental influences cause phenotypic differences between individuals. To this  
316 end, data from twins and their families (nearly 200,000 participants) from all over the  
317 Netherlands are collected, with a focus on health, lifestyle, personality, brain development,  
318 cognition, mental health, and aging. In NTR Biobank<sup>15</sup> samples for DNA, RNA, cell lines and  
319 for biomarker projects have been collected.

320 **RS**

321 The Rotterdam Study<sup>16</sup> is a single-centre, prospective population-based cohort study conducted  
322 in Rotterdam, the Netherlands<sup>16</sup>. Subjects were included in different phases, with a total of  
323 14,926 men and women aged 45 and over included as of late 2008. The main objective of the  
324 Rotterdam Study is to investigate the prevalence and incidence of and risk factors for chronic  
325 diseases to contribute to a better prevention and treatment of such diseases in the elderly.

326 **Genotype data**

327 ***Data generation***

328 Genotype data was generated for each cohort individually. Details on the methods used can be  
329 found in the individual papers (CODAM: van Dam et al.<sup>35</sup>; LLD: Tigchelaar et al.<sup>12</sup>; LLS:  
330 Deelen et al.<sup>39</sup>, 2014; NTR: Willemsen et al.<sup>15</sup>; RS: Hofman et al.<sup>16</sup>).

331 ***Imputation and QC***

332 For each cohort separately, the genotype data were harmonized towards the Genome of the  
333 Netherlands<sup>40</sup> (GoNL) using Genotype Hamonizer<sup>41</sup> and subsequently imputed per cohort using  
334 Impute2<sup>42</sup> using GoNL<sup>43</sup> reference panel<sup>43</sup> (v5). Quality control was also performed per cohort.  
335 We removed SNPs with an imputation info-score below 0.5, a HWE *P*-value smaller than  $10^{-4}$ , a  
336 call rate below 95% or a minor allele frequency smaller than 0.05. These imputation and filtering  
337 steps resulted in 5,206,562 SNPs that passed quality control in each of the datasets.

338 **Methylation data**

339 ***Data generation***

340 For the generation of genome-wide DNA methylation data, 500 ng of genomic DNA was  
341 bisulfite modified using the EZ DNA Methylation kit (Zymo Research, Irvine, California, USA)  
342 and hybridized on Illumina 450k arrays according to the manufacturer's protocols. The original  
343 IDAT files were generated by the Illumina iScan BeadChip scanner. We collected methylation  
344 data for a total of 3,841 samples. Data was generated by the Human Genotyping facility (HugeF)  
345 of ErasmusMC, the Netherlands ([www.glimDNA.org](http://www.glimDNA.org)).

346 ***Probe remapping and selection***

347 We remapped the 450K probes to the human genome reference (HG19) to correct for inaccurate  
348 mappings of probes and identify probes that mapped to multiple locations on the genome. Details  
349 on this procedure can be found in Bonder et al. (2014)<sup>44</sup>. Next, we removed probes with a known  
350 SNP (GoNL, MAF > 0.01) at the single base extension (SBE) site or CpG site. Lastly, we  
351 removed all probes on the sex chromosomes, leaving 405,709 high quality methylation probes  
352 for the analyses.

353 **Normalization and QC**

354 Methylation data was directly processed from IDAT files resulting from the Illumina 450k array  
355 analysis, using a custom pipeline based on the pipeline developed by Tost & Touleimat<sup>45</sup>. First,  
356 we used methylumi<sup>46</sup> to extract the data from the raw IDAT files. Next, we performed quality  
357 control checks on the probes and samples, starting by removing the incorrectly mapped probes.  
358 We checked for outlying samples using the first two principal components (PCs) obtained using  
359 principal component analysis (PCA). None of the samples failed our quality control checks,  
360 indicating high quality data. Following quality control, we performed background correction and  
361 probe type normalization as implemented in DASEN<sup>47</sup>. Normalization was performed per cohort,  
362 followed by quantile normalization on the combined data to normalize the differences per cohort.  
363 The next step in quality control consisted of identifying potential sample mix-ups between  
364 genotype and DNA methylation data. Using mix-up mapper<sup>48</sup>, we detected and corrected 193  
365 mix-ups. Lastly, in order to correct for known and unknown confounding sources of variation in  
366 the methylation data and to give us more power to detect meQTLs, we removed the first  
367 components which were not affected by genetic information, the 22 first PCs, from the  
368 methylation data using methodology we have successfully used in *trans*-eQTL<sup>3,49</sup> and meQTL  
369 analyses before<sup>44</sup>.

370 **RNA sequencing**

371 Total RNA from whole blood was deprived of globin using Ambion's GLOBIN clear kit and  
372 subsequently processed for sequencing using Illumina's Truseq version 2 library preparation kit.  
373 Paired-end sequencing of 2x50bp was performed using Illumina's Hiseq2000, pooling 10  
374 samples per lane. Finally, read sets per sample were generated using CASAVA, retaining only

375 reads passing Illumina's Chastity Filter for further processing. Data was generated by the  
376 Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands ([www.glimDNA.org](http://www.glimDNA.org)).

377 Initial QC was performed using FastQC<sup>50</sup> (v0.10.1), removal of adaptors was performed using  
378 cutadapt<sup>51</sup> (v1.1), and Sickle<sup>52</sup> (V1.2) [2] was used to trim low quality ends of the reads (min  
379 length 25, min quality 20). The sequencing reads were mapped to human genome (HG19) using  
380 STAR<sup>53</sup> v2.3.125 . Gene expression quantification was performed by HTseq-count. The gene  
381 definitions used for quantification were based on Ensmble version 71, with the extension that  
382 regions with overlapping exons were treated as separate genes and reads mapping within these  
383 overlapping parts did not count towards expression of the normal genes.

384 Expression data on the gene level were first normalized using Trimmed Mean of M-values<sup>54</sup>.  
385 Then expression values were log2 transformed, gene and sample means were centred to zero. To  
386 correct for batch effects, PCA was run on the sample correlation matrix and the first 25 PCs were  
387 removed using methodology that we have use for eQTL analyses before<sup>49,55</sup>. More details are  
388 provided in Zhernakova et al (in preperation).

### 389 ***Cis*-meQTL mapping**

390 In order to determine the effect of nearby genetic variation on methylation levels (*cis*-meQTLs),  
391 we performed *cis*-meQTL mapping using 3,841 samples for which both genotype data and  
392 methylation data were available. To this end, we calculated the Spearman rank correlation and  
393 corresponding *P*-value for each CpG-SNP pair in each cohort separately. We only considered  
394 CpG-SNP pairs located no further than 250kb apart. The *P*-values were subsequently  
395 transformed into a Z-score for meta-analysis. To maximize the power of meQTL detection, we  
396 performed a meta-analysis over all datasets by calculating an overall, joint *P*-value using a  
397 weighted Z-method. A comprehensive overview of this method has been described previously<sup>55</sup>.

398 To detect all possible independent SNPs regulating methylation at a single CpG-site we  
399 regressed out all primary *cis*-meQTL effects and then ran *cis*-meQTL mapping for the same  
400 CpG-site to find secondary *cis*-meQTL. We repeated that in a stepwise fashion until no more  
401 independent *cis*-meQTL were found.

402 To filter out potential false positive *cis*-meQTLs caused by SNPs affecting the binding of a probe  
403 on the array, we filtered the *cis*-meQTLs effects by removing any CpG-SNP pair for which the  
404 SNP was located in the probe. In addition, all other CpG-SNP pairs for which the SNP was  
405 outside the probe, but in LD ( $R^2 > 0.2$  or  $D' > 0.2$ ) with a SNP inside the probe were also  
406 removed. We tested for LD between SNPs in the probe and in the surrounding *cis* area in the  
407 individual genotype datasets, as well as in GoNL v5, in order to be as strict as possible in  
408 marking a QTL as true positive.

409 To correct for multiple testing, we empirically controlled the false discovery rate (FDR) at 5%.  
410 For this, we compared the distribution of observed *P*-values to the distribution obtained from  
411 performing the analysis on permuted data. Permutation was done by shuffling the sample  
412 identifiers of one data set, breaking the link between, e.g., the genotype data and the methylation  
413 or expression data. We repeated this procedure 10 times to obtain a stable distribution of *P*-  
414 values under the null distribution. The FDR was determined by only selecting the strongest effect  
415 per CpG<sup>55</sup> in both the real analysis and in the permutations (i.e. probe level FDR < 5%).

#### 416 ***Cis*-eQTL mapping**

417 For a set of 2,116 BIOS samples we had also generated RNA-seq data. We used this data to  
418 identify *cis*-eQTLs. *Cis*-eQTL mapping was performed using the same method as *cis*-meQTL  
419 mapping. Details on these eQTLs will be described in a separate paper (Zhernakova *et al*, in  
420 preparation).

421 **Expression quantitative trait methylation (eQTM) analysis**

422 To identify associations between methylation levels and expression levels of nearby genes (*cis*-  
423 eQTM)s), we first corrected our expression and methylation data for batch effects and covariates  
424 by regressing out the PCs and regressing out the identified *cis*-meQTLs and *cis*-eQTLs, to ensure  
425 only relationships between CpG sites and gene expression levels would be detected that were not  
426 attributable to particular genetic variation or batch effects. We mapped eQTM in a window of  
427 250Kb around the TSS of a transcript. Further statistical analysis was identical to the *cis*-meQTL  
428 mapping. For this analysis we were able to use a total of 2,101 samples for which both genetic,  
429 methylation and gene expression data was available. To correct for multiple testing we controlled  
430 the FDR at 5%, the FDR was determined by only selecting the strongest effect per CpG<sup>55</sup> in both  
431 the real analysis and in the permutations.

432 ***Trans*-meQTL mapping**

433 To identify the effects of distal genetic variation with methylation (*trans*-meQTLs) we used the  
434 same 3,841 samples that we had used for *cis*-meQTL mapping. To focus our analysis and limit  
435 the multiple testing burden, we restricted our analysis to SNPs that have been previously found  
436 to be significantly correlated to traits and diseases at a  $P < 5 \times 10^{-8}$ . We extracted these SNPs from  
437 the NHGRI genome-wide association study (GWAS) catalogue, used recent GWAS studies not  
438 yet in the NHGRI GWAS catalogue and studies on the Immunochip and Metabochip platform  
439 that are not included in the NHGRI GWAS catalogue (Extended Data Table 1). We compiled  
440 this list of SNPs in December 2014. Per SNP we only investigated CpG sites that mapped at least  
441 5 Mb from the SNP or on other chromosomes. Before mapping *trans*-meQTLs, we regressed out  
442 the identified *cis*-meQTLs to increase the statistical power of *trans*-meQTL detection (as done  
443 previously for *trans*-eQTLs<sup>3</sup>) and to avoid designating an association as *trans* that may be due to

444 long-range LD (e.g. within the HLA region). To ascertain the stability of the *trans*-meQTLs we  
445 also performed the *trans*-mapping on the non-corrected data and the methylation data corrected  
446 for cell-type proportions. In addition, we performed meQTL mapping on SNPs known to  
447 influence the cell type proportions in blood<sup>19,20</sup>.

448 To filter out potential false positive *trans*-meQTLs due to cross-hybridization of the probe, we  
449 remapped the methylation probes with very relaxed settings, identical to Westra et al.<sup>55</sup>, with the  
450 difference that we only accepted mappings if the last bases of the probe including the SBE site  
451 were mapped accurately to the alternative location. If the probe mapped within our minimal  
452 *trans*-window, 5 Mb from the SNP, we removed the effect as being a false positive *trans*-  
453 meQTL.

454 We controlled for multiple testing by using 10 permutations. We controlled the false-discovery  
455 rate at 5%, identical to the aforementioned *cis*-meQTL analysis.

#### 456 ***Trans*-eQTL mapping**

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

#### 462 **Annotations and enrichment tests**

463 Annotation of the CpGs was performed using Ensembl<sup>56</sup> (v70), UCSC Genome Browser<sup>57</sup> and  
464 data from the Epigenomics Roadmap Project.<sup>58</sup> We used the Epigenomics Roadmap annotation  
465 for the SBE site of the methylation site for all 27 blood cell types. We chose to use both the

466 histone mark information and the chromatin marks in blood-related cell types only, as generated  
467 by the Epigenomics Roadmap Project. Summarizing the information over the 27 blood cell types  
468 was done by counting presence of histone-marks in all the cell types and scaling the abundance,  
469 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  
470 blood cell types the score would be 0.

471 To calculate enrichment of meQTLs or eQTM<sub>s</sub> for any particular genomic context, we used  
472 logistic regression because this allows us to account for covariates such as CpG methylation  
473 variation. For *cis*-meQTLs, we used the variability of DNA methylation, the number of SNPs  
474 tested, and the distance to the nearest SNP per CpG as covariates. For all other analyses we used  
475 only the variability in DNA methylation as a covariate.

476 Next to annotation data from the Epigenomics Roadmap project, we used transcription factor  
477 ChIP-seq data from the ENCODE-project for blood-related cell lines. For every CpG site, we  
478 determined if there was an overlap with a ChIP-seq signal and performed a Fisher exact test to  
479 determine whether the *trans*-meQTL probes associated with the SNP in the transcription factor  
480 region of interest were more often overlapping with a ChIP-seq region than the other *trans*-  
481 meQTL probes. We collected all transcription factor called narrow peak files from the UCSC  
482 genome browser to perform the enrichments.

483 Enrichment of known sequence motifs among *trans*-CpGs was assessed by PWMEnrich<sup>22</sup>  
484 package in R, Homer<sup>59</sup> and DEEPbind<sup>23</sup>. For PWMEnrich hundred base pair sequences around  
485 the interrogated CpG site were used, and as a background set we used the top CpGs from the 50  
486 permutations used to determine the FDR threshold of the *trans*-meQTLs. For Homer the default  
487 settings for motif enrichment identification were used, and the same CpGs derived from the  
488 permutations were used as a background. For DEEPbind we used both the permutation

489 background like described for Homer and the permutations background as described for  
490 PWMEnrich.

491 Using data published by Rao *et al.*<sup>10</sup> we were able to intersect the *trans*-meQTLs with  
492 information about the 3D structure of the human genome. For the annotation, we used the  
493 combined Hi-C data for both inter- and intra-chromosomal data at 1Kb and the quality threshold  
494 of E30 in the GM12878 lymphoblastoid cell line. Both the *trans*-meQTL SNP and *trans*-meQTL  
495 probes were put in the relevant 1Kb block, and for these blocks we looked up the chromosomal  
496 contact value in the measurements by Rao et al. Surrounding the *trans*-meQTLs SNPs, we used a  
497 LD window that spans maximally 250Kb from the *trans*-meQTL SNP and had a minimal  $R^2$  of  
498 0.8. If a Hi-C contact between the SNP block and the CpG-site was indicated, we flagged the  
499 region as a positive for Hi-C contacts. As a background, we used the combinations found in our  
500 50 permuted *trans*-meQTL analyses, taking for each permutation the top *trans*-meQTLs that  
501 were similar in size to the real analysis. This permitted us to empirically determine whether there  
502 were significantly more Hi-C interactions in the real data as compared to the permutations.

503 **eQTM direction prediction**

504 We predicted the direction of the eQTM effects using both a decision tree and a naïve Bayes  
505 model (as implemented by Rapid-miner<sup>60</sup> v6.3). We built the models on the strongest eQTMs  
506 (i.e. those identified at a very stringent FDR  $<9.73 \times 10^{-6}$ ). For the decision tree we used a  
507 standard cross-validation set-up using 20 folds. For the naive Bayes model we used a double  
508 loop cross-validation: performance was evaluated in the outer loop using 20-fold cross-  
509 validation, while feature selection (using both backward elimination and forward selection) took  
510 place in the inner loop using 10-fold cross-validation. Details about the double-loop cross-  
511 validation can be found in Ronde et al.<sup>61</sup>. During the training of the model, we balanced the two

512 classes making sure we had an equal number of positively correlating and negatively correlating  
513 CpG-gene combinations, by randomly sampling a subset of the overrepresented negatively  
514 correlating CpG-gene combination group. We chose to do so to circumvent labelling al eQTMs  
515 as negative, since this is the class were the majority of the eQTMs are in.

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

## 519 **DEPICT**

520 To investigate whether there was biological coherence in the *trans*-meQTLs identified, we  
521 performed gene-set enrichment analysis for each genetic risk factor that was showing at least 10  
522 *trans*-meQTL effects. To do so, we adapted DEPICT<sup>27</sup>, a pathway enrichment analysis method  
523 that we previously developed for GWAS. Instead of defining loci with genes by using top  
524 associated SNPs, we used the eQTM information to link CpGs to genes. Within DEPICT gene  
525 set enrichment, significance is determined by using matched sets of permuted loci (in terms of  
526 numbers of genes per locus) that have been identified using simulated GWAS. Subsequent  
527 pathway enrichment analysis was conducted as described before, and significance was  
528 determined by controlling the false discovery rate at 5%.

## 529 **References**

530

- 531 1. Manolio, T. a. Genomewide association studies and assessment of the risk of disease. *N.  
532 Engl. J. Med.* **363**, 166–176 (2010).
- 533
- 534 2. Visscher, P. M., Brown, M. a., McCarthy, M. I. & Yang, J. Five years of GWAS  
535 discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

536

537 3. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known  
538 disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).

539

540 4. Wright, F. a *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat.*  
541 *Genet.* **46**, 430–7 (2014).

542

543 5. Bernstein, B. E., Meissner, A. & Lander, E. S. The Mammalian Epigenome. *Cell* **128**,  
544 669–681 (2007).

545

546 6. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with  
547 genetic variation in gene regulation. *Elife* **2**, e00523 (2013).

548

549 7. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell  
550 differentiation. *Nature* **518**, 344–349 (2015).

551

552 8. Yao, C. *et al.* Integromic analysis of genetic variation and gene expression identifies  
553 networks for cardiovascular disease phenotypes. *Circulation* **131**, 536–49 (2015).

554

555 9. Huan, T. *et al.* A Meta-analysis of Gene Expression Signatures of Blood Pressure and  
556 Hypertension. *PLOS Genet.* **11**, e1005035 (2015).

557

558 10. Rao, S. S. P., Huntley, M. H., Durand, N. C. & Stamenova, E. K. A 3D Map of the Human  
559 Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**,  
560 1665–1680 (2014).

561

562 11. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and  
563 Distal Chromosomal Interactions. *Cell* **162**, 1051–65 (2015).

564

565 12. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population  
566 cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ*  
567 *Open* **5**, e006772 (2015).

568

569 13. van Greevenbroek, M. M. J. *et al.* The cross-sectional association between insulin  
570 resistance and circulating complement C3 is partly explained by plasma alanine  
571 aminotransferase, independent of central obesity and general inflammation (the CODAM  
572 study). *Eur. J. Clin. Invest.* **41**, 372–379 (2011).

573

574 14. Schoenmaker, M. *et al.* Evidence of genetic enrichment for exceptional survival using a  
575 family approach: the Leiden Longevity Study. *Eur. J. Hum. Genet.* **14**, 79–84 (2006).

576

577 15. Willemsen, G. *et al.* The Adult Netherlands Twin Register: twenty-five years of survey  
578 and biological data collection. *Twin Res. Hum. Genet.* **16**, 271–81 (2013).

579

580 16. Hofman, A. *et al.* The rotterdam study: 2014 objectives and design update. *Eur. J.*  
581 *Epidemiol.* **28**, 889–926 (2013).

582

583 17. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription  
584 factors. *Elife* **2013**, 1–16 (2013).

585

586 18. Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic variation  
587 located at thousands of independent loci. *Nat. Commun.* **6**, 6326 (2015).

588

589 19. Orrù, V. *et al.* Genetic variants regulating immune cell levels in health and disease. *Cell*  
590 **155**, 242–56 (2013).

591

592 20. Roederer, M. *et al.* The Genetic Architecture of the Human Immune System: A  
593 Bioresource for Autoimmunity and Disease Pathogenesis. *Cell* **161**, 387–403 (2015).

594

595 21. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime  
596 cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**,  
597 576–589 (2010).

598

599 22. Stojnic, R. & Diez, D. PWMEnrich: PWM Enrichment Analysis.

600

601 23. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence  
602 specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**,  
603 831–838 (2015).

604

605 24. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene  
606 expression in human cells. *Proc. Natl. Acad. Sci.* **111**, 996–1001 (2013).

607

608 25. Splinter, E. *et al.* CTCF mediates long-range chromatin looping and local histone  
609 modification in the beta-globin locus. *Genes Dev.* **20**, 2349–54 (2006).

610

611 26. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of

612        inflammatory bowel disease. *Nature* **491**, 119–24 (2012).

613

614    27. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using  
615        predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).

616

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

620

621    29. Lettre, G. *et al.* Genome-Wide association study of coronary heart disease and its risk  
622        factors in 8,090 african americans: The nhlbi CARe project. *PLoS Genet.* **7**, (2011).

623

624    30. Soranzo, N. *et al.* Meta-analysis of genome-wide scans for human adult stature identifies  
625        novel loci and associations with measures of skeletal frame size. *PLoS Genet.* **5**, (2009).

626

627    31. Filion, G. J. P. *et al.* A Family of Human Zinc Finger Proteins That Bind Methylated  
628        DNA and Repress Transcription A Family of Human Zinc Finger Proteins That Bind  
629        Methylated DNA and Repress Transcription. *Mol. Cell. Biol.* **26**, 169 (2006).

630

631    32. Sasai, N. & Defossez, P. A. Many paths to one goal? The proteins that recognize  
632        methylated DNA in eukaryotes. *Int. J. Dev. Biol.* **53**, 323–334 (2009).

633

634    33. Shiraishi, K. *et al.* A genome-wide association study identifies two new susceptibility loci  
635        for lung adenocarcinoma in the Japanese population. *Nat. Genet.* **44**, 900–903 (2012).

636

637    34. Qiu, Z. *et al.* Functional Interactions between NURF and Ctcf Regulate Gene Expression.  
638        *Mol. Cell. Biol.* **35**, 224–237 (2015).

639

640    35. Van Dam, R. M., Boer, J. M. a, Feskens, E. J. M. & Seidell, J. C. Parental history off  
641        diabetes modifies the association between abdominal adiposity and hyperglycemia.  
642        *Diabetes Care* **24**, 1454–1459 (2001).

643

644    36. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank.  
645        *Int. J. Epidemiol.* 1–9 (2014). doi:10.1093/ije/dyu229

646

647    37. Boomsma, D. I. *et al.* Netherlands Twin Register: a focus on longitudinal research. *Twin*  
648        *Res.* **5**, 401–406 (2002).

649

650 38. Boomsma, D. I. *et al.* Genome-wide association of major depression: description of  
651 samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank  
652 projects. *Eur. J. Hum. Genet.* **16**, 335–342 (2008).

653

654 39. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a  
655 novel locus conferring survival beyond 90 years of age. *Hum. Mol. Genet.* **23**, 4420–4432  
656 (2014).

657

658 40. The Genome of the Netherlands Consortium. Whole-genome sequence variation,  
659 population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 1–  
660 95 (2014).

661

662 41. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion  
663 for genotype data integration. *BMC Res. Notes* **7**, 901 (2014).

664

665 42. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation  
666 method for the next generation of genome-wide association studies. *PLoS Genet.* **5**,  
667 (2009).

668

669 43. Deelen, P. *et al.* Improved imputation quality of low-frequency and rare variants in  
670 European samples using the ‘Genome of The Netherlands’. *Eur. J. Hum. Genet.* 1–6  
671 (2014). doi:10.1038/ejhg.2014.19

672

673 44. Bonder, M. J. *et al.* Genetic and epigenetic regulation of gene expression in fetal and adult  
674 human livers. *BMC Genomics* **15**, 860 (2014).

675

676 45. Touleimat, N. T echnology R eport Complete pipeline for Infinium ® Human Methylation  
677 450K BeadChip data processing using subset quantile normalization for accurate DNA  
678 methylation estimation T echnology R eport. *4*, 325–341 (2012).

679

680 46. Davis, S., Du, P., Bilke, S., Triche, T. J. & Bootwalla, M. methylumi: Handle Illumina  
681 methylation data.

682

683 47. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation  
684 array data. *BMC Genomics* **14**, 293 (2013).

685

686 48. Westra, H.-J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets  
687 increases power to detect small genetic effects. *Bioinformatics* **27**, 2104–11 (2011).

688

689 49. Fehrmann, R. S. N. *et al.* Trans-eQTLs reveal that independent genetic variants associated  
690 with a complex phenotype converge on intermediate genes, with a major role for the HLA.  
691 *PLoS Genet.* **7**, e1002197 (2011).

692

693 50. FastQC. at <<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>>

694

695 51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.  
696 **17**, 10–12 (2011).

697

698 52. Joshi, N. A. & Fass, J. N. Sickle: A sliding-window, adaptive, quality-based trimming tool  
699 for FastQ files. (2011). at <<https://github.com/najoshi/sickle>>

700

701 53. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21  
702 (2013).

703

704 54. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential  
705 expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

706

707 55. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known  
708 disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).

709

710 56. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, 48–55 (2013).

711

712 57. Kent, W. J. *et al.* The Human Genome Browser at UCSC The Human Genome Browser at  
713 UCSC. *Genome Res.* 996–1006 (2002). doi:10.1101/gr.229102.

714

715 58. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature*  
716 **518**, 317–330 (2015).

717

718 59. Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function.  
719 *Nature* **503**, 487–92 (2013).

720

721 60. Markus, Hofmann Klinkenberg, R. *RapidMiner: Data Mining Use Cases and Business*  
722 *Analytics Applications*. (Chapman & Hall/CRC, 2014).

723

724 61. de Ronde, J. J., Bonder, M. J., Lips, E. H., Rodenhuis, S. & Wessels, L. F. a. Breast cancer  
725 subtype specific classifiers of response to neoadjuvant chemotherapy do not outperform  
726 classifiers trained on all subtypes. *PLoS One* **9**, e88551 (2014).

727

728

729 **Acknowledgements**

730 This work was performed within the framework of the Biobank-Based Integrative Omics Studies  
731 (BIOS) Consortium funded by BBMRI-NL, a research infrastructure financed by the Dutch  
732 government (NWO 184.021.007). Samples were contributed by LifeLines  
733 (<http://lifelines.nl/lifelines-research/general>), the Leiden Longevity Study ([http://www.healthy-](http://www.healthy-ageing.nl)  
734 [ageing.nl](http://www.leidenlangleven.nl); <http://www.leidenlangleven.nl>), the Netherlands Twin Registry  
735 (NTR: <http://www.tweelingenregister.org>), the Rotterdam studies, ([http://www.erasmus-](http://www.erasmus-epidemiology.nl/rotterdamstudy)  
736 [epidemiology.nl/rotterdamstudy](http://www.erasmus-epidemiology.nl/rotterdamstudy)), the Genetic Research in Isolated Populations program  
737 (<http://www.epib.nl/research/geneticepi/research.html#gip>), the Codam study  
738 (<http://www.carimmaastricht.nl/>) and the PAN study (<http://www.alsonderzoek.nl/>). We thank the  
739 participants of all aforementioned biobanks and acknowledge the contributions of the investigators  
740 to this study (Supplemental Acknowledgements). This work was carried out on the Dutch national  
741 e-infrastructure with the support of SURF Cooperative.

742

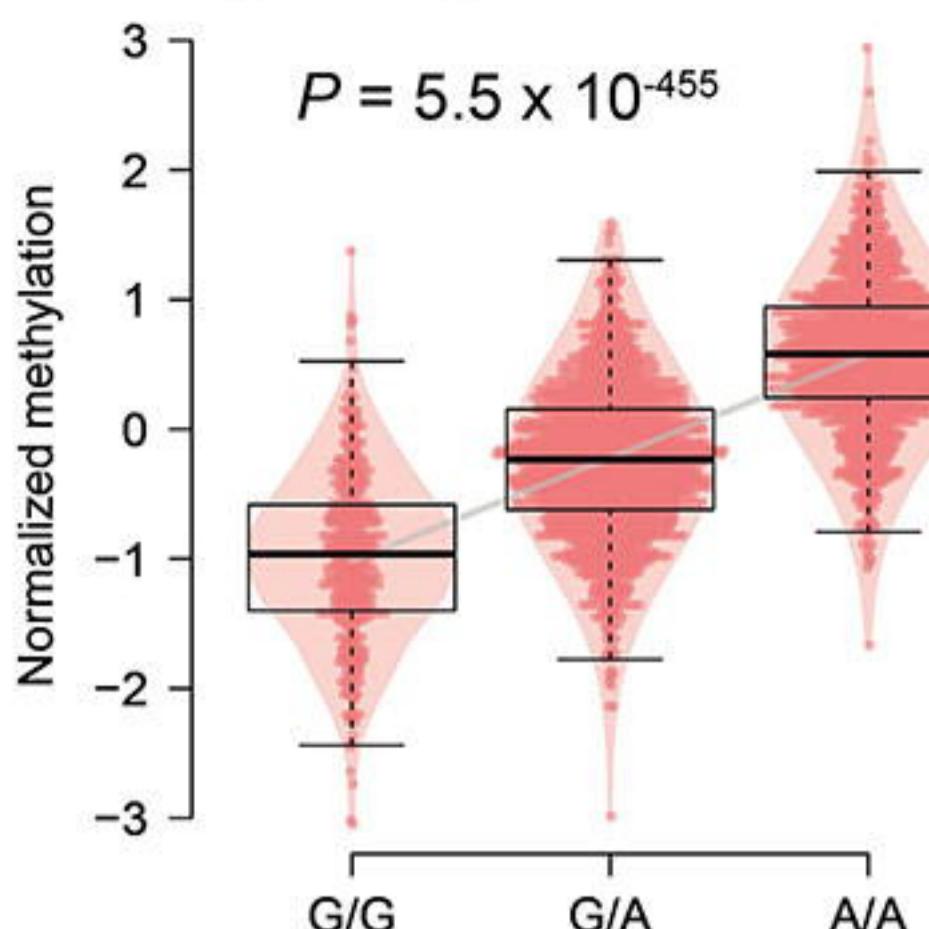
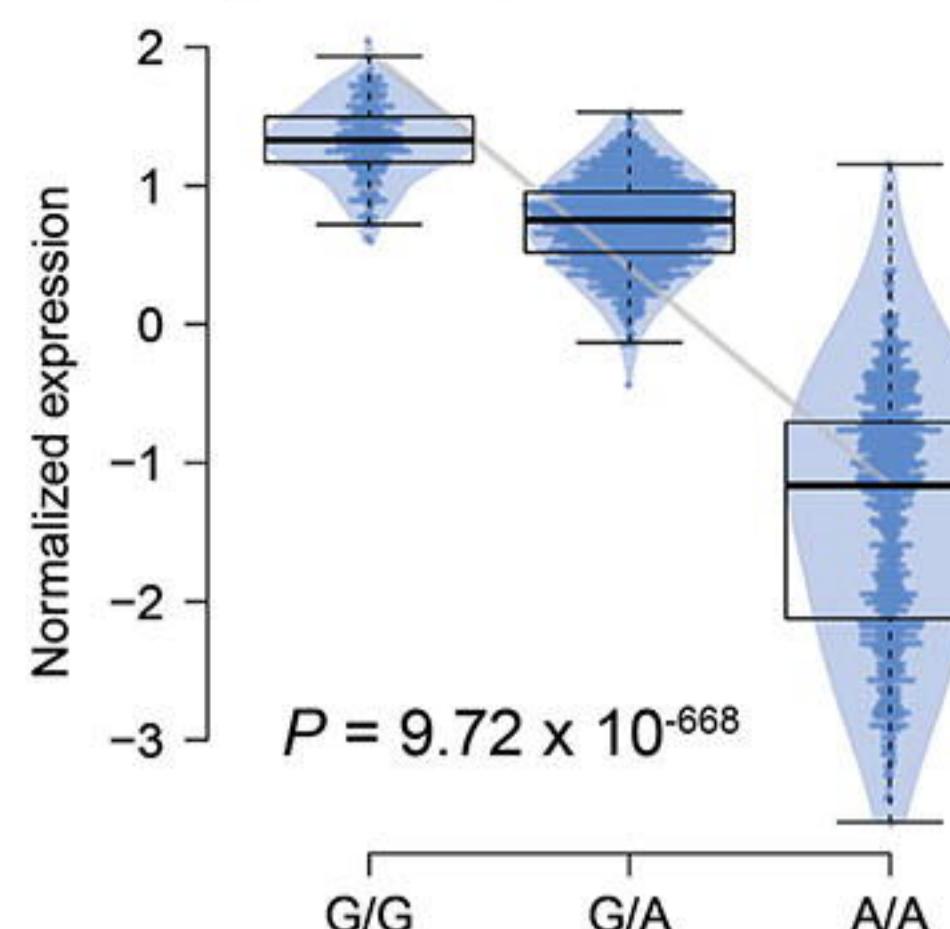
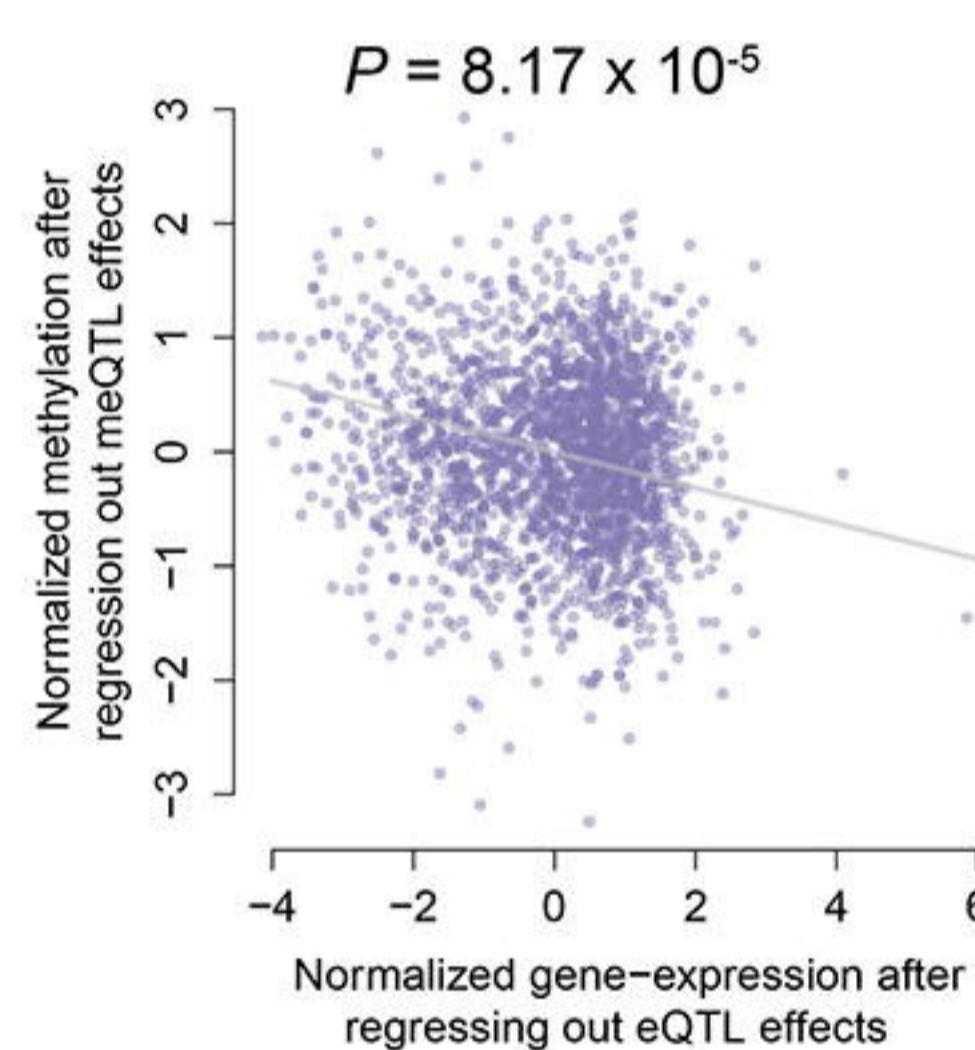
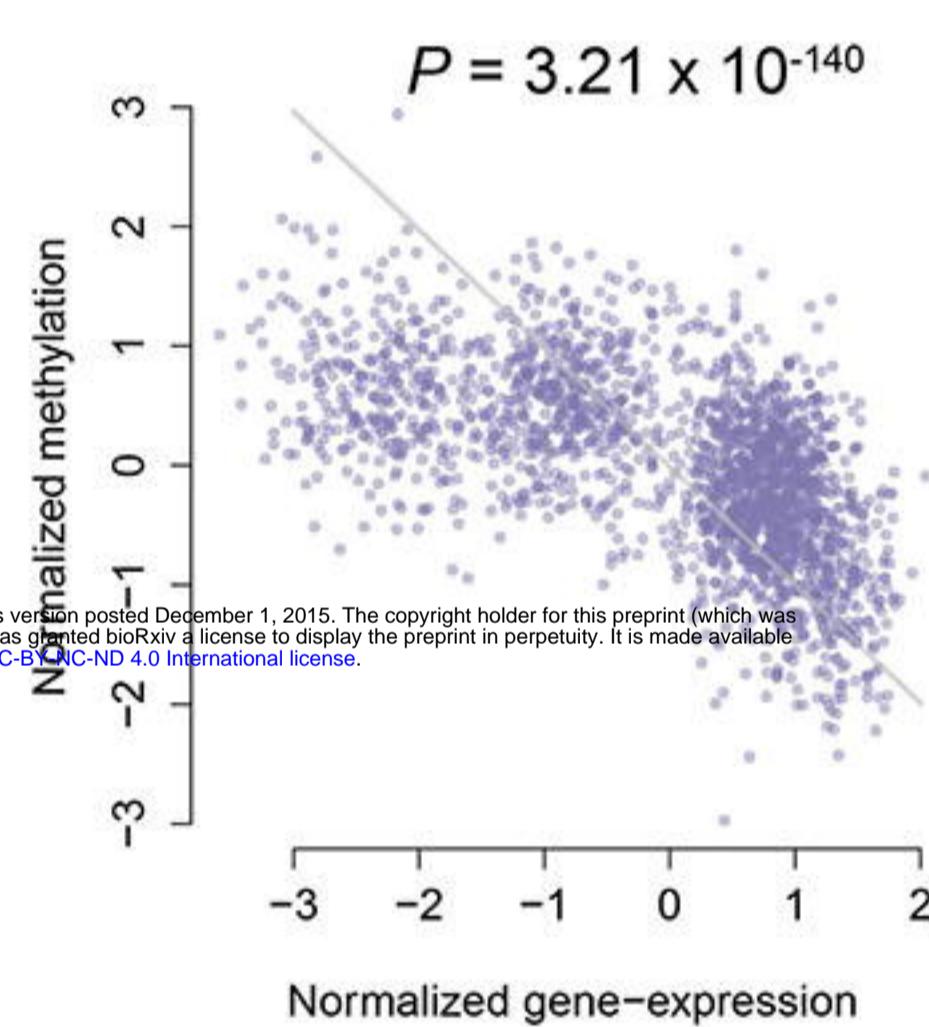
743

744 **Author contributions**

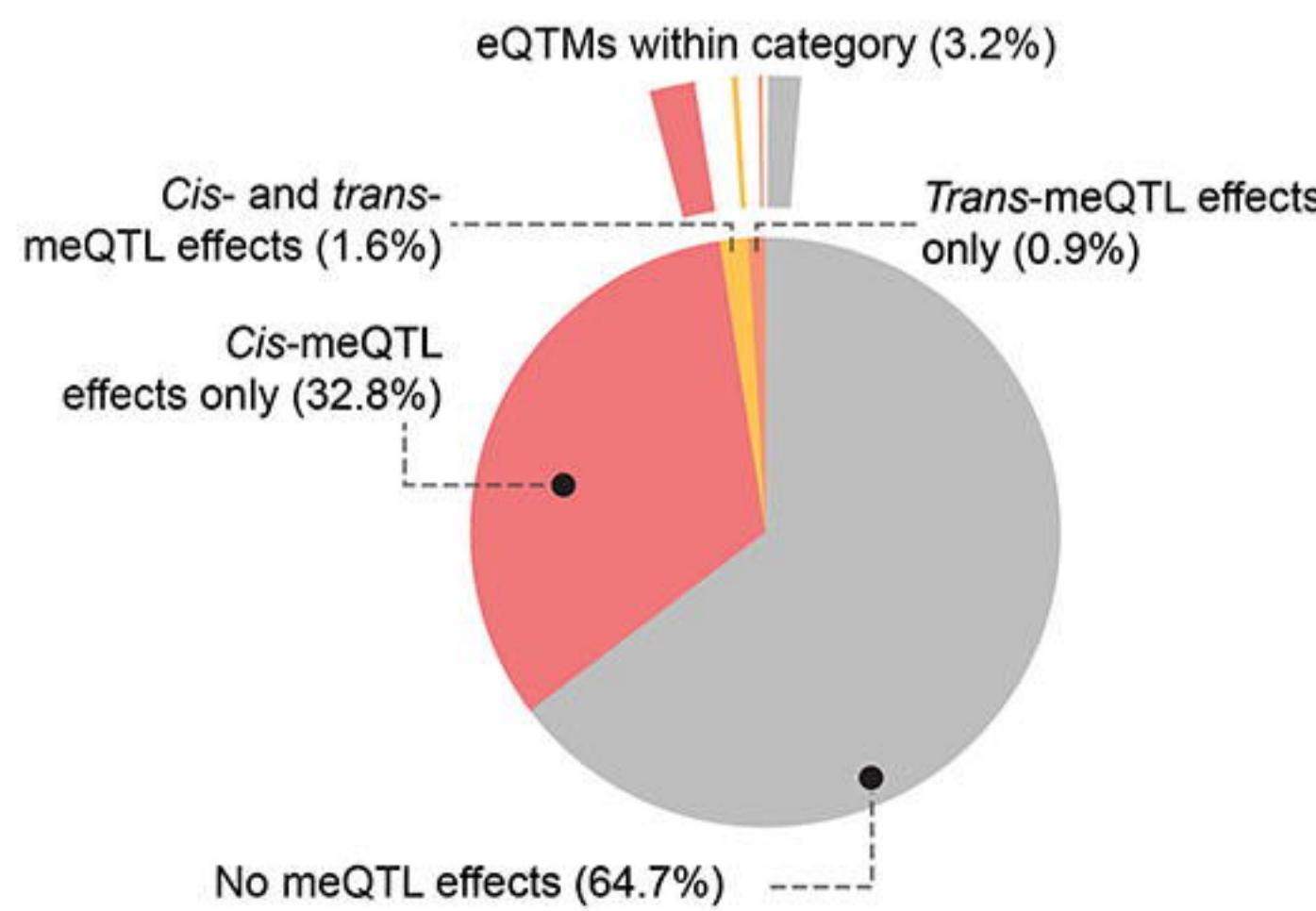
745 BTH, PACtH, JBJvM, AI, RJ and LF formed the management team of the BIOS consortium. DIB,  
746 RP, JVD, JJH, MMJVG, CDAS, CJHvdK, CGS, CW, LF, AZ, EFG, PES, MB, JD, DvH, JHV,  
747 LHvdB, CMvD, BAH, AI, AGU managed and organized the biobanks. JBJvM, PMJ, MV, HEDS,  
748 MV, RvdB, JvR and NL generated RNA-seq and Illumina 450k data. HM, MvI, MvG, JB, DVZ,  
749 RJ, PvtH, PD, IN, PACtH, BTH and MM were responsible for data management and the  
750 computational infrastructure. MJB, RL, MV, DVZ, RS, IJ, MvI, PD, FvD, MvG, WA, SMK, MAS,  
751 EWvZ, RJ, PACtH, LF and BTH performed the data analysis. MJB, RL, LF and BTH drafted the  
752 manuscript.

753 **Data availability**

754 All results can be queried using our dedicated QTL browser:  
755 <http://genenetwork.nl/biosqlbrowser/>. Raw data was submitted to the European Genome-  
756 phenome Archive (EGA).

**a Methylation Quantitative Trait Loci****b Expression Quantitative Trait Loci****meQTL****eQTL****c****Expression Quantitative Trait Methylation**

bioRxiv preprint doi: <https://doi.org/10.1101/033084>; this version posted December 1, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**d****Many CpG-sites are under genetic control****e****Trans-meQTLs are robust across cell types**