HMG Advance Access published July 28, 2015

Layered genetic control of DNA methylation and gene expression: a locus of multiple sclerosis in healthy individuals

Jean Shin[1,13], Celine Bourdon[1,13], Manon Bernard[1], Michael Wilson[1], Eva Reischl[2], Melanie Waldenberger[2], Barbara Ruggeri[3], Gunter Schumann[3], Sylvane Desrivieres[3], Alexander Leemans[4], the IMAGEN Consortium, the SYS Consortium, Michal Abrahamowicz[5], Gabriel Leonard[6], Louis Richer[7], Luigi Bouchard[8,9], Daniel Gaudet[9,10], Tomas Paus[11,12], and Zdenka Pausova[1,*]

[1]The Hospital for Sick Children, University of Toronto, Toronto, Canada

[2]Research Unit of Molecular Epidemiology, Helmholtz Zentrum Munchen, Munich, Germany

[3]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[4]Image Sciences Institute, University Medical Center Utrecht, Utrecht, the Netherlands

[5]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada

[6]Montreal Neurological Institute and Hospital, McGill University, Montreal, Canada

[7]Department of Psychology, Université du Québec à Chicoutimi, Chicoutimi, Canada

[8]Department of Biochemistry, Université de Sherbrooke, Sherbrooke, Canada

[9]ECOGENE-21 and Lipid Clinic, Chicoutimi Hospital, Chicoutimi, Canada

[10]Department of Medicine, Université de Montréal, Montréal, Canada

[11]Rotman Research Institute, University of Toronto, Toronto, Canada

[12]Child Mind Institute, New York, NY, USA

[*]The author for correspondence: Zdenka Pausova, MD, FAHA, Senior Scientist, The Hospital for Sick Children, Associate Professor, Departments of Physiology and Nutritional Sciences, University of Toronto, Toronto, Canada, Phone: (416) 813-7654/4340; Fax: (416) 813-5771, E-mail: zdenka.pausova@sickkids.ca

[13]Authors with equal contribution

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

ABSTRACT

DNA methylation may contribute to the etiology of complex genetic disorders through its impact on genome integrity and gene expression; it is modulated by DNA-sequence variants, named *methylation quantitative-trait loci* (meQTLs). Most meQTLs influence methylation of a few CpG dinucleotides within short genomic regions (<3kb). Here we identified a layered genetic control of DNA methylation at numerous CpGs across a long 300-kb genomic region. This control involved a single *long-range* meQTL and multiple *local* meQTLs. The *long-range* meQTL explained up to 75% of variance in methylation of CpGs located over extended areas of the 300-kb region. The meQTL was identified in four samples ($p=2.8x10^{-17}$, $p=3.1x10^{-31}$, $4.0x10^{-71}$, $5.2x10^{-199}$), comprising a total of 2,796 individuals. The *long-range* meQTL was strongly associated not only with DNA methylation but also with mRNA expression of several genes within the 300-kb region ($p=7.1x10^{-18}$–$1.0x10^{-123}$). The associations of the meQTL with gene expression became attenuated when adjusted for DNA methylation (causal inference test: $p=2.4\times10^{-13}$–7.1×10^{-20}), indicating coordinated regulation of DNA methylation and gene expression. Further, the *long-range* meQTL was found to be in linkage disequilibrium with the most replicated locus of multiple sclerosis, a disease affecting primarily the brain white matter. In middle-aged adults free of the disease, we observed that the risk allele was associated with subtle structural properties of the brain white matter found in multiple sclerosis ($p=0.02$). In summary, we identified a *long-range* meQTL that controls methylation and expression of several genes and may be involved in increasing brain vulnerability to multiple sclerosis.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

INTRODUCTION

DNA methylation (DNAm) – the addition of a methyl group at the fifth position of cytosines in CpG dinucleotides (CpGs) – is one of the most studied epigenetic modifications (1). It has been implicated in the regulation of genome integrity and gene expression and, as such, has the potential to be involved in the etiology of complex genetic disorders (2, 3).

In DNA samples extracted from multiple cells, DNAm is a quantitative trait, measured as a proportion of DNA strands that are methylated. DNAm is modulated by DNA-sequence variants, termed *methylation quantitative trait loci* (meQTLs) (4-10). Some meQTLs impact methylation of one or a few CpGs, whereas others influence methylation of multiple CpGs distributed most often across short (<3kb) and – occasionally – also across long (>3kb) genomic segments (9). Thus, methylation status of an individual CpG is likely the result of a combined effect of *local* (within CpGs), *short-range* (<3kb) and *long-range* (>3kb) meQTLs. This 'layered genetic control' of DNAm has not been studied extensively.

Here we focused on the 1,000 most variable CpGs as assessed with the Illumina 450K BeadChip interrogating >450,000 CpGs across the genome (11). Through a subset of these highly variable CpGs, we uncovered a layered genetic control of DNAm across a long (300-kb) genomic region that involved a single *long-range* meQTL and multiple *local* meQTLs. The *long-range* meQTL was strongly associated not only with DNAm but also mRNA expression of genes within the 300-kb region. It was found to be in linkage disequilibrium with the most replicated locus of multiple sclerosis (MS) (12-15), a disease of the brain that affects white matter. Furthermore, in a population-based subsample of middle-aged adults, we observed that the allele of the *long-range* meQTL enhancing risk for MS was associated with subtle variations in structural properties of the brain white matter similar to those found in non-lesional white matter in patients with MS (16). Thus, we identified a *long-range* meQTL that controls methylation and expression of multiple genes and may be involved in increasing brain vulnerability to MS. We made these observations in four samples from three independent cohorts: (*i*) 132 adolescents from the Saguenay Youth Study (SYS) study (17, 18), (*ii*) 278 parents from the SYS study (17, 18), (*iii*) 639 adolescents from the IMAGEN study (19) and (*iv*) 1,747 participants from the Ontario Familial Colon Cancer Registry (OFCCR) Study (20, 21)(Table S1).

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

RESULTS

Identification of a 'long-range' meQTL

We discovered a *long-range* meQTL as follows. First, as part of a routine quality control, we performed principal component analysis (PCA) of the 1,000 most variable CpGs in the genome. This initial PCA was carried out in a sample of 132 SYS adolescents using DNAm β values adjusted for age, sex, batch and blood cell fractions (22). The largest component of shared variance – PC1 (Figure S1) – was loaded by 25 CpGs distributed across a 300-kb segment of chromosome 6, which included a total of eight genes (Table 1). PC1 explained 40% of the variance shared among the 25 CpGs. Similar results were observed in three other samples (278 SYS parents, 639 IMAGEN adolescents and 1,747 OFCCR participants; Table 1). This cross-sample similarity suggested that PC1, and its loading by specific CpGs, may be determined genetically.

Next, we estimated heritability of DNAm at the 25 highly variable CpGs loading into PC1. These analyses showed that DNAm is highly heritable at these CpGs, with the heritability estimates being up to 0.90 (Figure S2). To identify the specific genetic factors determining PC1, we performed a genome-wide association study (GWAS) of PC1 in the discovery sample of 132 SYS adolescents. This analysis identified a single locus (rs4959030, $p=2.8x10^{-17}$) residing within an inter-genic region between *HLA-DRB1* and *HLA-DQA1* (Figure 1). The same locus was also found in SYS parents ($p=3.1x10^{-31}$, Figure 1) and was replicated in IMAGEN ($4.0x10^{-71}$) and OFCCR ($p=5.2x10^{-199}$, Table 2). Thus, we identified a *long-range* meQTL – a DNA-sequence variant that was associated with shared variance in DNAm among 25 CpGs distributed across a large genomic region (300kb).

Next, we determined the contribution of this *long-range* meQTL to methylation of each individual CpG loading into PC1. Nineteen of these 25 CpGs were polymorphic, *i.e.*, known single nucleotide polymorphisms were located within these CpGs (23); therefore, we also considered the influences of these *local* meQTLs. We used multivariate models that assessed, at each CpG, the relative contributions of the *long-range* meQTL and respective *local* meQTL (when present). With this approach, we observed that the *long-range* meQTL, independent of respective *local* meQTLs, mainly contributed to methylation of CpGs within five of the above eight genes contained in the 300-kb segment of chromosome 6. Relative to the *long-range* meQTL, three of these genes were located 'telomerically' (*HLA-DRB5*, *HLA-DRB6* and *HLA-*

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

DRB1) and two were located 'centromerically' (*HLA-DQA1* and *HLA-DQB1*) (Figure 2). The contribution was greater for the 'telomerically' located than 'centromerically' located genes – the *long-range* meQTL explained up to 75% of total variance in DNAm within the 'telomerically located' genes and up to 25% of total variance in DNAm within the 'centromerically located'' genes (Figure 2). At the 19 polymorphic CpGs, the respective *local* meQTLs, independent of the *long-range* meQTL, explained 5-30% of the variance (Figures 2 and S3). These results indicate that the *long-range* meQTL plays an important role in DNAm of the five *HLA-DR* and *HLA-DQ* genes.

The above analyses were performed using only the 1,000 most variable CpGs in the genome. Next, we examined whether the *long-range* meQTL effects extend to less variable CpGs within the 300-kb region (an additional 431 CpGs [including 66 polymorphic] assessed with the Illumina 450K chip). This exploration revealed that, indeed, the *long-range* meQTL was also associated with less variable CpGs within the 300-kb region (Figure 2).

Genomic landscape of the long-range meQTL

The mechanisms of how *long-range* meQTLs modulate DNAm are not well understood. They may influence DNAm through their impact on transcription-factor and chromatin-modifier binding and subsequent alterations in chromatin structure and accessibility to DNAm machinery (9). Therefore, we examined the genomic landscape of the studied 300-kb region with ENCODE (24).

The *long-range* meQTL that we identified was located in an inter-genic region between *HLA-DRB1* and *HLA-DQA1* – approximately 29kb from *HLA-DRB1* and 15kb from *HLA-DQA1* (Figure 2). Within this region, it was positioned within an area of DNaseI hypersensitivity, indicating an open chromatin state (25). This area was also enriched for features of active enhancers (26), namely, high H3K4me1 and H3K27Ac signals and low H3K4me3 signal (Figure 3). Further, it contained binding sites for a large group of transcriptional factors, cofactors, chromatin regulators and transcription apparatus, which is characteristic of so-called super-enhancers (27) (Figure 3). Super-enhancers are large clusters of transcriptional enhancers that usually drive expression of genes important for defining cell identity during development (27). Compared with typical enhancers, super-enhancers are larger in size and transcription-factor density, they have a greater ability to activate transcription and they are more sensitive to

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

perturbations, such as reduced levels of enhancer-bound factors and co-factors (28). Based on the ENCODE ChIP-seq data in the GM12878 lymphoblastoid cell line (24), we observed that the super-enhancer containing the *long-range* meQTL we identified here includes genomic regions bound by (*i*) PU.1 and PAX5 (transcription factors acting as master regulators of myeloid and lymphoid differentiation (29, 30)), (*ii*) CTCF and cohesin complex (chromatin regulators), and (*iii*) RNA polymerase II, TBF and TAF1 (key factors and cofactors of the transcription apparatus, Figure 3).

The *long-range* meQTL that we identified lies within a eutherian DNA-repetitive element (MamRep1879) that is primate-specific (Figure S4). To examine whether it alters transcription-factor binding motif predictions, we used the regulatory variant prediction software HaploReg v2(31). HaploReg made eight predictions (Table S3), with the most striking one being the presence of a PU.1 transcription-factor binding site (PU.1_disc3) with the major allele and its absence with the minor allele (Figure S4B). This PU.1 motif fell within an existing PU.1 ChIP-seq peak (chr6: 32,591,515-32,591,757) identified in the GM12878 lymphoblastoid cell line, which is homozygous for the major allele (24). The predicted PU.1 motif (chr6: 32,591,749-32,591,758) occurs 109 bp downstream of the best matching PU.1 motif. It remains to be seen whether the loss of the second PU.1 motif by the minor allele of the *long-range* meQTL will affect PU.1 binding at the region.

Taken together, the above genomic-landscape information suggests the possibility that the *long-range* meQTL, being located within a super-enhancer, may regulate not only DNAm but also mRNA expression of neighboring genes by modulating enhancer-promoter interactions; this may occur via CTCF homodimerisation and associated chromatin looping (32-34). Therefore, we examined whether the meQTL was associated not only with DNAm of the five genes but also with their mRNA expression.

The 'long-range' meQTL and mRNA expression of neighboring genes

One of the five neighboring genes was a pseudogene (albeit transcribed, *HLA-DRB6*) and, as such, was not assayed by the employed expression chip and not examined in the present study. The analysis of the remaining four genes showed that mRNA expression of three of them was associated with the *long-range* meQTL. Specifically, the minor allele of the meQTL was strongly associated with higher expression of *HLA-DRB5* and *HLA-DRB1* ($p=1.0x10^{-123}$ and $p=3.5x10^{-48}$,

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

respectively) and less strongly with lower expression of *HLA-DQB1* ($p=7.1x10^{-18}$, Table 3 and Figure 4). The meQTL explained a total of 48%, 26% and 13% of the variance in mRNA expression of *HLA-DRB5*, *HLA-DRB1*, and *HLA-DQB1*, respectively.

Next, we assessed to what extent each of the associations between the *long-range* meQTL and mRNA levels were dependent on the association between the meQTL and DNAm (PC1). This analysis tested the possibility that the meQTL impacts DNAm and mRNA expression through a shared molecular pathway altered by the same event. We hypothesized this event would be a meQTL-induced change in transcription-factor binding within the super-enhancer affecting local chromatin structure, gene expression and DNAm. These analyses showed that the associations between the meQTL and mRNA expression became significantly attenuated when additionally adjusted for DNAm (PC1, Table 3). The proportion of variance in mRNA expression explained by the meQTL decreased from 48% to 29% for *HLA-DRB5*, from 26% to 15% for *HLA-DRB1*, and from 13% to 5% for *HLA-DQB1*. Based on the causal inference test (35), these decreases were significant at $p=7.1\times10^{-20}$, 2.4×10^{-13} and 7.5×10^{-17}, respectively (Table 3). These results suggest that DNAm and mRNA expression of these three genes are in part co-regulated, with the co-regulator being the identified *long-range* meQTL.

The 'long-range' meQTL, multiple sclerosis and structural properties of the brain white matter

The *long-range* meQTL that we identified is located within the major histocompatibility complex class II (MHC-II) region on chromosome 6p21. The genes within this region, including the *DR* and *DQ* genes studied here, encode antigen-presenting molecules. The MHC-II region has been associated with a number of diseases related to immune dysregulation. Thus, we assessed whether the *long-range* meQTL is in linkage disequilibrium with any disease loci mapped previously to the region of this meQTL (Figure 5). This analysis showed that the *long-range* meQTL is in linkage disequilibrium ($r^2=0.80$) with the most replicated and strongest ($p=10^{-225}$) locus of multiple sclerosis (MS, Figure 5 and Table S4) (12-15). The top SNPs of this locus tag the classical HLA allele *DRB1*1501* (13, 36, 37). Loci of other autoimmune disorders, such as rheumatoid arthritis and systemic sclerosis, have also been mapped to this genomic region, but they were not in linkage disequilibrium with this *long-range* meQTL (Table S4).

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

Multiple sclerosis is an autoimmune disease affecting primarily the brain white matter (38). Pre-symptomatic white-matter degeneration may be part of the pathogenesis of MS (39). One imaging marker of MS is lower fractional anisotropy (FA) of white matter (16). As assessed with diffusion tensor imaging, FA of white matter depends on the microstructural features of fiber tracts, including the relative alignment of individual axons, their packing density, myelin content and axon caliber (40). In patients with MS, FA is lower within lesions than within normal-appearing white-matter tissue, and when assessed within normal-appearing white-matter tissue, FA is lower in MS patients than in healthy controls (16). In the SYS cohort, we collected data on FA of the brain white matter in a subset of 309 middle-aged adults free of MS. In this dataset, we observed that the *long-range* meQTL allele linked previously to higher risk for MS was associated here with lower FA ($p=0.02$, Figure S5).

DISCUSSION

In the present study, we identified a layered genetic control of DNAm at numerous CpGs across a long 300-kb MHC-II region. This control involves a single *long-range* meQTL and multiple *local* meQTLs. The *long-range* meQTL explains up to 75% of variance in methylation of CpGs located over extended areas of the 300-kb region. This *long-range* meQTL regulates not only methylation but also expression of several genes. It is in linkage disequilibrium with the major locus of MS (12-15). In the present study, the allele associated with higher risk for MS was associated with markedly higher expression of *HLA-DRB5* and *HLA-DRB1* and modestly lower expression of *HLA-DQB1*. Finally, the meQTL risk-allele was associated with lower FA in the brain white matter in a subsample of middle-aged individuals.

The *long-range* meQTL that we identified is located intergenically – between *HLA-DRB1* and *HLA-DQA1* – within a super-enhancer (27). Super-enhancers are large clusters of transcriptional enhancers that drive expression of genes important for defining cell identity and functionality (27). Our results demonstrate that the meQTL impacts both DNAm and mRNA expression. We hypothesize it does so through processes triggered by the same molecular event. This event may be a meQTL-induced change in transcription-factor binding of the super-enhancer that alters super-enhancer/promoter interactions and, in turn, chromatin structure, mRNA expression and DNAm. Compared with typical enhancers, super-enhancers are more sensitive to perturbations, such as reduced levels of bound transcription factors (28). The minor

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

allele of the identified meQTL is predicted to abolish a PU.1 motif and, as such, may reduce the super-enhancer's binding by this transcription factor. This change in turn may alter the super-enhancer's capacity to interact with neighboring promoters and drive mRNA expression from these promoters. Thus, depending on the meQTL allele, the super-enhancer may interact with either the *HLA-DRB1* promoter (minor allele) or the *HLA-DQB1* promoter (major allele). When interacting with the *HLA-DRB1* promoter (minor allele), it activates mRNA expression of *HLA-DRB1* and two neighboring genes transcribed in the same direction (*HLA-DRB6* and *HLA-DRB5*) (Figure 4). When interacting with the *HLA-DQB1* promoter (major allele), it creates a chromatin loop and activate mRNA expression of *HLA-DQB1* but not that of the neighboring *HLA-DQA1* (as this gene is transcribed in the opposite direction). This scenario is consistent with the mRNA-expression results observed in the present study – the minor allele of the *long-range* meQTL was associated with markedly *higher* expression of *HLA-DRB1* and *HLA-DRB5*, modestly *lower* expression of *HLA-DQB1* and *no* effect on expression of *HLA-DQA1* (Figure 4). Overall, as all these genes encode antigen-presenting molecules, the minor allele of this *long-range* meQTL may be associated with aberrant and possibly augmented antigen-presenting activity.

It has been reported recently that gene expression and DNAm may be regulated in a coordinate fashion by the same genetic variants (9). Consistent with this possibility, we showed here that the observed associations between the *long-range* meQTL and expression of each *HLA-DRB1*, *HLA-DRB5* and *HLA-DQB1* became attenuated when adjusted for DNAm (PC1). The proportion of variance explained by the meQTL decreased from 48% to 29% for *HLA-DRB5*, from 26% to 15% for *HLA-DRB1*, and from 13% to 5% for *HLA-DQB1*, and these decreases were significant as tested with the causal inference test (35).

In the present study, the *long-range* meQTL showed lower contributions to both methylation and mRNA expression of the 'centromeric' (*HLA-DQA1* and *HLA-DQB1*) vs. 'telomeric' (*HLA-DRB5*, *HLA-DRB6* and *HLA-DRB1*) genes. Whether these differences relate to DNA looping, which we propose involves only the 'centromeric' group of genes (Figure 4), requires further experimental research. This possibility is supported by previous research suggesting that chromatin remodelling (such as DNA looping) alters the accessibility of DNA to methylation and transcription machineries and thus DNA methylation and transcription within the involved genomic region (41, 42).

Here we found that the identified *long-range* meQTL is in linkage disequilibrium with the

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

major locus of MS (12-15). MS is an autoimmune disease affecting predominantly the brain white matter. The most common form of MS is characterized by recurring episodes of inflammatory demyelination and progressive neurodegeneration. MS is thought to emerge in genetically susceptible individuals when they encounter environmental triggers that initiate an inflammatory reaction against self-antigens in the brain (13, 38). Some argue that the primary pathogenic process of MS is neurodegeneration associated with excess myelin debris that (being strongly antigenic) triggers immune reaction (39).

The identified *long-range* meQTL may explain some of the genetic susceptibility for MS; it may also add to our understanding of possible underlying mechanisms. The minor allele of this meQTL, which is in linkage disequilibrium with the risk allele for MS (12-15), is predicted to abolish a PU.1 disc3 motif. PU.1 (also known as *SPI1*) is a master transcription factor critical for myeloid hematopoiesis (29), as well as for the development and function of microglia, the main resident immune cells of the brain (43-45). Unlike other cells in the brain, microglia share their developmental origin with blood cells – they are derived from embryonic hematopoietic precursors that seed the brain prior to birth (46). Microglia are surveyors of the brain microenvironment searching constantly for areas of local injury and homeostasis disturbances (47). In most brain diseases, microglia engulf pathogens, dead cells, myelin debris and misfolded proteins (47). Recently, it has been suggested that microglia play similar homeostatic roles in the 'healthy' brain (48). In performing these roles, microglia undergo graded 'activation'. During some 'activation' states, microglia mount immune reactions and begin expressing MHC-II genes (including those studied here) (47). Whether the minor allele of the *long-range* meQTL associated in the present study with overexpression of certain MHC-II genes modulates this capacity of microglia, requires further research.

In the present study, the minor allele of the *long-range* meQTL we identified was associated with lower FA in the brain white matter of middle-aged adults free of MS. FA of white matter, as assessed with diffusion tensor imaging, depends on the microstructural features of fiber tracts, including the relative alignment of individual axons, their packing density, myelin content and axon caliber (40). Patients with MS commonly exhibit diffusely abnormal white matter in which myelin and axonal volumes are reduced but inflammation is not apparent (49). In MS patients, FA is lower within lesions than within normal-appearing white-matter tissue (16). Importantly, FA is lower in MS patients than in healthy controls even within normal-appearing

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

tissue (16). Whether the observed FA differences associated with the *long-range* meQTL genotypes relate to the diffuse white matter abnormalities seen in MS (49) and the proposed pre-symptomatic neurodegeneration (39) requires further research.

In this study, the *long-range* meQTL was similarly associated with DNAm in adolescent and adult samples (Table 2), suggesting the association is not impacted by age. Further, the association was identified in DNA extracted from peripheral blood cells. A considerable similarity in DNAm exists across different tissues; bisulfite sequencing of DNA from 12 different human tissues showed that only 5-15% of CpGs are methylated in a tissue-specific manner (50). The degree of between–tissue similarity is higher for developmentally close (vs. distant) tissues (50, 51). Thus, given the shared developmental origin of peripheral blood cells and brain microglia (46), DNAm patterns may be highly similar in the two cell types. This possibility requires further research.

DNAm is a complex quantitative trait. Twin and family-based studies suggest it is influenced by multiple genetic and environmental factors (4, 5). In the present study, we estimated heritability of DNAm at all CpGs assessed within the studied 300-kb genomic region (Figure S2). The results showed that, at numerous CpGs, genetic factors explained up to 90% of total variance. At many of these CpGs, the identified *long-range* meQTL explained up to 75% of total variance. The remaining variance could be explained by *local* meQTLs (up to 25%) and/or other – yet unidentified – genetic and environmental factors.

In summary, we identified a *long-range* meQTL that coordinates methylation and expression of several genes within the MHC-II region and may increase vulnerability to MS.

MATERIALS AND METHODS

Cohorts

The present study was conducted in four samples from three cohorts. The three cohorts are (*i*) the SYS Study (separate samples of adolescents and parents), (*ii*) the IMAGEN Study, and (*iii*) the OFCCR Study. The basic characteristics of these cohorts are provided in Table S1. Written consents of adults and assent of adolescents (and consent of their parents) were obtained. The regional research ethics committees approved the study protocols.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

The SYS Study

Recruitment: The Saguenay Youth Study (SYS) is a population-based cross-sectional study of cardio-metabolic and brain health in adolescents and their parents (n=1,028 adolescents and 949 parents). The cohort was recruited from the genetic founder population of the Saguenay Lac St. Jean region of Quebec, Canada, via adolescents. Additional recruitment details and selection criteria have been described previously (17, 18). The present study was conducted on a subset of the SYS adolescents (n=132) and parents (n=278) on whom genome–wide genotyping and epityping (described below) have been conducted.

Epityping was conducted on DNA extracted from peripheral blood cells using the Infinium HumanMethylation450K BeadChip (Illumina, San Diego, CA) at the Helmholtz Zentrum München German Research Center for Environmental Health (Neuherberg, Germany) in adolescents and at the Montreal Genome Centre (Montreal, Canada) in parents. The Infinium HumanMethylation450K BeadChip interrogates methylation at >485,000 CpGs (52). The DNAm score at each CpG, i.e., the DNAm β value, is derived from the fluorescent intensity ratio [β=intensity of the methylated allele/(intensity of the unmethylated allele + intensity of the methylated allele + 100)] (11). Adolescent and parent samples were randomly loaded onto 11 and 28 arrays (12 samples per array), respectively. DNAm β values were normalized using the Subset-Quantile Within Array Normalization (SWAN) procedure (53). Quality control was performed by excluding CpGs with detection $p \geq 0.05$ in more than 20% of samples (764 CpGs in adolescents and 912 CpGs in parents). After excluding these probes, as well as control probes and probes on sex chromosomes, a total of 473,608 CpGs were analyzed. All samples had >98% sites with detection $p<0.05$.

Genotyping: The SYS adolescents and parents were genotyped in two waves. First, 592 adolescents were genotyped with the Illumina Human610-Quad BeadChip (Illumina, San Diego, CA; n=582,892 SNPs) at the Centre National de Génotypage (Paris, France). Second, the remaining 427 adolescents and all parents were genotyped with the HumanOmniExpress BeadChip (Illumina, San Diego, CA; n=729,295 SNPs) at the Genome Analysis Centre of Helmholtz Zentrum München (Munich, Germany). In both genotyping waves, SNPs with call rate <95% and minor allele frequency <0.01, and SNPs that were not in Hardy-Weinberg equilibrium ($p<1x10^{-6}$) were excluded. After this quality control, 542,345 SNPs on the first chip and 644,283 SNPs on the second chip were available for analysis.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

Genotype imputation was used to equate the set of SNPs genotyped on each platform and to increase the SNP density. Haplotype phasing was performed with SHAPEIT (54) using an overlapping subset of 313,653 post-quality-control SNPs that were present on both genotyping platforms and the 1,000 Genomes SNPs in European reference panel (Phase 1, Release 3). Imputation was conducted on the phased data with IMPUTE2 (55). Markers with low imputation quality (information score <0.5) or low minor allele frequency (<0.01) were removed. After this quality control of imputation, a total of 7,746,837 typed and imputed SNPs were analyzed.

Diffusion tensor imaging of the brain was available in a subset of parents. It was conducted as follows. All brain imaging data were acquired with a 1.5 Tesla Siemens (Avanto) scanner. The T1-weighted images were acquired using the 3D Magnetization Prepared Rapid Gradient Echo sequence with 176 sagittal slices (1-mm isotropic resolution, TR = 2,400ms, TE = 2.65ms, TI = 1000ms, and flip angle = 8°). The Diffusion Tensor Images were acquired in two runs using echo planar scans with 61 axial slices, 2-mm isotropic resolution, TR = 8,000ms, TE = 94ms and flip angle=90°; 64 diffusion directions were acquired with a b-value of 1000 s/mm^2. ExploreDTI (56) was used to correct for eddy current-induced geometric distortions and motion. In addition, the participant's T1-weighted image was used to apply a correction for echo planar imaging and susceptibility distortions (57). Diffusion tensors were estimated at each voxel using the RESTORE method for robust estimation of tensors by outlier rejection (58). For each participant, the average fractional anisotropy (FA) values were obtained for all white-mater voxels contained in the four lobes; the lobe masks were projected to each participant's native DTI space using a non-linear registration. All steps of this processing pipeline were quality controlled (e.g., no movement artifacts, correct registration). Here, we analyzed mean FA across the four lobes in 309 parents with quality-controlled data.

The IMAGEN Study

Recruitment: The IMAGEN Study is a European multi-center study on impulsivity, reinforcement sensitivity and emotional reactivity in adolescents (n=2,000; aged 13 to 15 years) (19). The IMAGEN Study is a community-based study; participants were recruited in local schools at eight participating sites in Germany (Berlin, Dresden, Hamburg, Manheim), United Kingdom (London and Nottingham), France (Paris) and Ireland (Dublin) between 2008 and 2010. The current study was conducted on a randomly selected subset of the IMAGEN participants in

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

whom genome-wide gene expression was assessed (n=639). All participants and their parents provided informed written assent and consent, respectively (19).

Genotyping was carried out with the Illumina Human610-Quad BeadChip (Illumina, San Diego, CA) at the Centre National de Génotypage (Paris, France). SNPs with call rate <95% and minor allele frequency <0.01, and SNPs that were not in Hardy-Weinberg equilibrium ($p<1x10^{-6}$) were excluded. After this quality control, 97 SNPs were available for analysis.

Epityping was conducted on DNA extracted from peripheral blood cells using the Infinium HumanMethylation450K BeadChip (Illumina, San Diego, CA) at the SNP&SEQ Technology Platform, Uppsala University (Uppsala, Sweden). DNAm β values were normalized using Genome Studio. Quality control was performed by excluding CpGs with detection p<0.01; 25 CpGs were available for the replication analysis. All samples had >98% sites with detection p<0.01.

Gene-expression was conducted on RNA extracted from fresh peripheral blood cells (collected into the PAXgene blood RNA tubes [Qiagen]) using the Illumina HumanHT-12 v4 Expression BeadChip (Illumina, San Diego, CA). Data were normalized using the *mloess* method(59). Only genes within the studied 300-kb MHC-II region (n=8) were analyzed in the present study.

The OFCCR Study

Recruitment: The Ontario Familial Colon Cancer Registry (OFCCR) sample is a population-based sample of colorectal cancer patients and their families (20, 21). The sample analyzed here consisted of 891 cases and 856 healthy controls who had genome-wide genetic and epigenetic data.

Genotyping: The OFCCR individuals were genotyped as described elsewhere(60). In brief, data from the Illumina 1,536 GoldenGate array, the Affymetrix/ParAllele 10K coding-SNP array and the Affymetrix Human Mapping 100K array were combined and complemented with the Affymetrix Human Mapping 500K array.

Epityping was conducted using the Infinium HumanMethylation450K BeadChip (Illumina, San Diego, CA) on DNA extracted from lymphocytes. Lymphocyte pellets were extracted from whole blood using Ficoll-Paque PLUS (GE Healthcare). DNA was extracted from lymphocytes using phenol-chloroform or Qiagen Mini-Amp DNA kit.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

Statistical analyses

First, we conducted *principal component analysis* (PCA) of 1,000 most variable CpGs in the genome. This analysis was part of a routine quality-control procedure aimed at identifying potentially confounding factors. We performed this PCA in the discovery sample of 132 SYS adolescents using DNAm β values adjusted for age, sex, batch and blood cell fractions (22). The top principal component (PC1, Figure S1) was loaded (>0.4) exclusively by CpGs (n=25) from a 300-kb segment of the MHC-II region on chromosome 6. We then performed replication PCAs with these 25 CpGs in a sample of 278 SYS parents and in two independent samples (639 IMAGEN adolescents and 1,747 OFCCR participants). The high consistency of the PC-based clusters of DNAm we observed across the four samples suggested genetic control of this clustering.

Second, to search for genes underlying the observed clustering of DNAm across the 25 CpGs loading into PC1, we conducted a *genome-wide association study* (GWAS) of PC1. This GWAS was carried out with the ProbABEL software (61) in the discovery sample of SYS adolescents (n=132). We then performed *replication studies* in the SYS parents (n=278), IMAGEN adolescents (n=639) and OFCCR participants (n=1,747). These studies were performed with all available SNPs within the studied 300-kb region of chromosome 6. Depending on the format of available genotypes, GWAS was conducted with either the ProbABEL software (61) (SYS parents, estimated genotypes, score test) or R (62) (IMAGEN and OFCCR participants, observed genotypes, Kruskal-Wallis test). In SYS adolescents and parents, PC1 was quantile-normalized using a rank-based inverse normal transformation prior to these association analyses (63).

Third, in the discovery sample of SYS adolescents, we determined the *relative contributions* of the GWAS-identified SNP (the *long-range* meQTL [rs4959030]) to DNAm variation at each of the 25 CpGs. As 19 of these 25 CpGs were polymorphic (i.e., known SNPs were located within these CpGs (23) [Table S2]), we also considered the contribution of these *local* meQTLs (minor allele frequency ≥5%). This was achieved by fitting, at each of the 25 CpGs, a linear regression model that included the GWAS-identified *long-range* meQTL and 1 or 2 respective *local* meQTLs onto age-, sex-, batch- and blood-cell fractions-adjusted DNAm β values. From this model, the relative contributions (partial R^2s) of the *long-range* meQTL and the

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

local meQTL(s) were determined by averaging sequential sums of squares over all possible orderings of the regressors. These analyses were also performed with all assessed CpGs within the studied 300-kb region (additional 472 CpGs). All calculations were conducted with the R package *relaimpo* (64).

Fourth, to define the *genomic landscape* of the GWAS-identified *long-range* meQTL and associated CpGs, we used the ENCODE database (ENCyclopedia Of DNA Elements, Hg19, http://genome.ucsc.edu/)(24). With this tool, we mapped features of chromatin structure, histone modifications, binding sites of transcription factors at promoter and enhancer elements, and regions of transcription within the studied 300-kb region.

Fifth, we tested whether the *long-range* meQTL [rs4959030]) associated with DNAm of eight neighboring genes was also associated with *mRNA expression* of these genes. We also tested whether these SNP associations with mRNA expression levels were independent of the SNP associations with DNAm. These latter analyses were performed using the causal inference test (35), involving quantile-normalized mRNA-expression levels, quantile-normalized PC1 and additively coded genotypes, while adjusting for sex and imaging center. The gene-expression analyses were conducted in the only sample with gene-expression data (IMAGEN, n=639).

Sixth, we examined whether the *long-range* meQTL (rs4959030) was in *linkage disequilibrium* with disease-oriented loci previously mapped to the region of this meQTL (chromosome 6:31.5-32.9Mb). To find these loci, we used the National Human Genome Research Institute (NHGRI) GWAS Catalog, which is a curated resource of SNP-trait associations (http://www.genome.gov/gwastudies/). Linkage disequilibrium was assessed with PLINK (65) using genotypes from the 1,000 Genomes Project European reference sample (March 2012).

Finally, we tested whether the *long-range* meQTL (rs4959030) was associated with fractional anisotropy, which is a structural property of the brain white matter that is frequently found to be lower in multiple sclerosis (16). Tests of association were performed using linear regression, assuming additive genetic model, while adjusting for age and sex. Prior to the association tests, fractional anisotropy values were quantile-normalized, using rank-based inverse normal transformation. These analyses were carried out in SYS parents (n=309).

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

ACKNOWLEDGEMENTS

We thank Helene Simard MA and her team of research assistants (Cégep de Jonquière) for their contributions in acquiring data for the SYS. The Canadian Institutes of Health Research and the Heart and Stroke Foundation of Canada fund the SYS. The McLaughlin Centre at the University of Toronto provided supplementary funds for the DNA methylation studies in the SYS. We thank Melissa Pangelinan and Deborah Schwartz for their help with analyzing DTI images.

We would like to thank Drs. Mathieu Lemire and Thomas Hudson for providing epigenetic and genetic data collected in the Ontario Familial Colon Cancer Registry.

CONFLICT OF INTEREST STATEMENT

None declared.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

REFERENCES

1 Rakyan, V.K., Down, T.A., Balding, D.J. and Beck, S. (2011) Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, **12**, 529-541.

2 Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M. *et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.*, **31**, 142-147.

3 Weisenberger, D.J. (2014) Characterizing DNA methylation alterations from The Cancer Genome Atlas. *J. Clin. Invest.*, **124**, 17-23.

4 Bjornsson, H.T., Sigurdsson, M.I., Fallin, M.D., Irizarry, R.A., Aspelund, T., Cui, H., Yu, W., Rongione, M.A., Ekstrom, T.J., Harris, T.B. *et al.* (2008) Intra-individual change over time in DNA methylation with familial clustering. *JAMA*, **299**, 2877-2883.

5 Kaminsky, Z.A., Tang, T., Wang, S.C., Ptak, C., Oh, G.H., Wong, A.H., Feldcamp, L.A., Virtanen, C., Halfvarson, J., Tysk, C. *et al.* (2009) DNA methylation profiles in monozygotic and dizygotic twins. *Nat. Genet.*, **41**, 240-245.

6 Liu, Y., Li, X., Aryee, M.J., Ekstrom, T.J., Padyukov, L., Klareskog, L., Vandiver, A., Moore, A.Z., Tanaka, T., Ferrucci, L. *et al.* (2014) GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. *Am. J. Hum. Genet.*, **94**, 485-495.

7 Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y. and Pritchard, J.K. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, **12**, R10.

8 Numata, S., Ye, T., Hyde, T.M., Guitart-Navarro, X., Tao, R., Wininger, M., Colantuoni, C., Weinberger, D.R., Kleinman, J.E. and Lipska, B.K. (2012) DNA methylation signatures in development and aging of the human prefrontal cortex. *Am. J. Hum. Genet.*, **90**, 260-272.

9 Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K. and Gilad, Y. (2014) Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.*, **10**, e1004663.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

10 Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J. *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.*, **6**, e1000952.

11 Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J.M., Delano, D., Zhang, L., Schroth, G.P., Gunderson, K.L. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288-295.

12 De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I., Ottoboni, L., Aggarwal, N.T., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C. *et al.* (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.*, **41**, 776-782.

13 Patsopoulos, N.A., Bayer Pharma, M.S.G.W.G., Steering Committees of Studies Evaluating, I.-b., a, C.C.R.A., Consortium, A.N., GeneMsa, International Multiple Sclerosis Genetics, C., Esposito, F., Reischl, J., Lehr, S. *et al.* (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann. Neurol.*, **70**, 897-912.

14 Australia and New Zealand Multiple Sclerosis Genetics, C. (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat. Genet.*, **41**, 824-828.

15 International Multiple Sclerosis Genetics, C., Hafler, D.A., Compston, A., Sawcer, S., Lander, E.S., Daly, M.J., De Jager, P.L., de Bakker, P.I., Gabriel, S.B., Mirel, D.B. *et al.* (2007) Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.*, **357**, 851-862.

16 Fox, R.J. (2008) Picturing multiple sclerosis: conventional and diffusion tensor imaging. *Semin. Neurol.*, **28**, 453-466.

17 Pausova, Z., Paus, T., Abrahamowicz, M., Almerigi, J., Arbour, N., Bernard, M., Gaudet, D., Hanzalek, P., Hamet, P., Evans, A.C. *et al.* (2007) Genes, maternal smoking, and the offspring brain and body during adolescence: design of the Saguenay Youth Study. *Hum. Brain Mapp.*, **28**, 502-518.

18 Paus, T., Pausova, Z., Abrahamowicz, M., Gaudet, D., Leonard, G., Pike, G.B. and Richer, L. (2014) Saguenay Youth Study: A multi-generational approach to studying virtual trajectories of the brain and cardio-metabolic health. *Dev. Cogn. Neurosci.*, in press.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

19 Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Buchel, C., Conrod, P.J., Dalley, J.W., Flor, H., Gallinat, J. *et al.* (2010) The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry*, **15**, 1128-1139.

20 Cotterchio, M., McKeown-Eyssen, G., Sutherland, H., Buchan, G., Aronson, M., Easson, A.M., Macey, J., Holowaty, E. and Gallinger, S. (2000) Ontario familial colon cancer registry: methods and first-year response rates. *Chronic Dis. Can.*, **21**, 81-86.

21 Newcomb, P.A., Baron, J., Cotterchio, M., Gallinger, S., Grove, J., Haile, R., Hall, D., Hopper, J.L., Jass, J., Le Marchand, L. *et al.* (2007) Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol. Biomarkers Prev.*, **16**, 2331-2343.

22 Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K. and Kelsey, K.T. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**, 86.

23 Chen, Y.A., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J. and Weksberg, R. (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, **8**, 203-209.

24 ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

25 Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E. *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390-394.

26 Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311-318.

27 Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A. and Young, R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934-947.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

28 Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I. and Young, R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307-319.

29 Rosenbauer, F. and Tenen, D.G. (2007) Transcription factors in myeloid development: balancing differentiation with transformation. *Nat. Rev. Immunol.*, **7**, 105-117.

30 Cobaleda, C., Schebesta, A., Delogu, A. and Busslinger, M. (2007) Pax5: the guardian of B cell identity and function. *Nat. Immunol.*, **8**, 463-470.

31 Ward, L.D. and Kellis, M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucl. Acids Res.*, **40**, D930-D934.

32 Ong, C.T. and Corces, V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, **15**, 234-246.

33 Majumder, P., Gomez, J.A., Chadwick, B.P. and Boss, J.M. (2008) The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. *J. Exp. Med.*, **205**, 785-798.

34 Majumder, P. and Boss, J.M. (2010) CTCF controls expression and chromatin architecture of the human major histocompatibility complex class II locus. *Mol. Cell. Biol.*, **30**, 4211-4223.

35 Millstein, J., Zhang, B., Zhu, J. and Schadt, E.E. (2009) Disentangling molecular relationships with a causal inference test. *BMC Genet.*, **10**, 23.

36 Sawcer, S., Franklin, R.J. and Ban, M. (2014) Multiple sclerosis genetics. *Lancet. Neurol.*, **13**, 700-709.

37 de Bakker, P.I., McVean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., Delgado, M. *et al.* (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.*, **38**, 1166-1172.

38 Ciccarelli, O., Barkhof, F., Bodini, B., De Stefano, N., Golay, X., Nicolay, K., Pelletier, D., Pouwels, P.J., Smith, S.A., Wheeler-Kingshott, C.A. *et al.* (2014) Pathogenesis of multiple sclerosis: insights from molecular and metabolic imaging. *Lancet. Neurol.*, **13**, 807-822.

39 Stys, P.K., Zamponi, G.W., van Minnen, J. and Geurts, J.J. (2012) Will the real multiple sclerosis please stand up? *Nat. Rev. Neurosci.*, **13**, 507-514.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

40 Beaulieu, C. (2002) The basis of anisotropic water diffusion in the nervous system - a technical review. *NMR Biomed.*, **15**, 435-455.

41 Portela, A. and Esteller, M. (2010) Epigenetic modifications and human disease. *Nat. Biotech.*, **28**, 1057-1068.

42 Jjingo, D., Conley, A.B., Yi, S.V., Lunyak, V.V. and Jordan, I.K. (2012) On the presence and role of human gene-body DNA methylation. *Oncotarget*, **3**, 462-474.

43 McKercher, S.R., Torbett, B.E., Anderson, K.L., Henkel, G.W., Vestal, D.J., Baribault, H., Klemsz, M., Feeney, A.J., Wu, G.E., Paige, C.J. *et al.* (1996) Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *EMBO J.*, **15**, 5647-5658.

44 Kierdorf, K., Erny, D., Goldmann, T., Sander, V., Schulz, C., Perdiguero, E.G., Wieghofer, P., Heinrich, A., Riemke, P., Holscher, C. *et al.* (2013) Microglia emerge from erythromyeloid precursors via Pu.1- and Irf8-dependent pathways. *Nat. Neurosci.*, **16**, 273-280.

45 Smith, A.M., Gibbons, H.M., Oldfield, R.L., Bergin, P.M., Mee, E.W., Faull, R.L. and Dragunow, M. (2013) The transcription factor PU.1 is critical for viability and function of human brain microglia. *Glia*, **61**, 929-942.

46 Ginhoux, F., Lim, S., Hoeffel, G., Low, D. and Huber, T. (2013) Origin and differentiation of microglia. *Front. Cell. Neurosci.*, **7**, 45.

47 Graeber, M.B. (2010) Changing face of microglia. *Science*, **330**, 783-788.

48 Hughes, V. (2012) Microglia: The constant gardeners. *Nature*, **485**, 570-572.

49 Seewann, A., Vrenken, H., van der Valk, P., Blezer, E.L., Knol, D.L., Castelijns, J.A., Polman, C.H., Pouwels, P.J., Barkhof, F. and Geurts, J.J. (2009) Diffusely abnormal white matter in chronic multiple sclerosis: Imaging and histopathologic analysis. *Arch. Neurol.*, **66**, 601-609.

50 Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378-1385.

51 Pai, A.A., Bell, J.T., Marioni, J.C., Pritchard, J.K. and Gilad, Y. (2011) A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.*, **7**, e1001316.

52 Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M. and Esteller, M. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, **6**, 692-702.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

53 Maksimovic, J., Gordon, L. and Oshlack, A. (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome biology*, **13**, R44.

54 Delaneau, O., Zagury, J.F. and Marchini, J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Meth.*, **10**, 5-6.

55 Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

56 Leemans, A., Jeurissen, B., Sijbers, J. and Jones, D. (2009) ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. *Proc. Intl. Soc. Mag. Reson. Med.*, **17**, 3537.

57 Irfanoglu, M.O., Walker, L., Sarlls, J., Marenco, S. and Pierpaoli, C. (2012) Effects of image distortions originating from susceptibility variations and concomitant fields on diffusion MRI tractography results. *NeuroImage*, **61**, 275-288.

58 Chang, L.C., Jones, D.K. and Pierpaoli, C. (2005) RESTORE: Robust estimation of tensors by outlier rejection. *Magn. Reson. Med.*, **53**, 1088-1095.

59 Šášik, R., Woelk, C.H. and Corbeil, J. (2004) Microarray truths and consequences. *J. Mol. Endocrinol.*, **33**, 1-9.

60 Zanke, B.W., Greenwood, C.M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S.M., Prendergast, J., Olschwang, S., Chiang, T., Crowdy, E. *et al.* (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **39**, 989-994.

61 Aulchenko, Y.S., Struchalin, M.V. and van Duijn, C.M. (2010) ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, **11**, 134.

62 RCoreTeam. (2014), R Foundation for Statistical Computing, Vienna, Austria, Vol. 2014.

63 Servin, B. and Stephens, M. (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.*, **3**, e114.

64 Grömping, U. (2006) Relative Importance for Linear Regression in R: The Package relaimpo. *J. Stat. Softw.*, **17**, 1-27.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

65 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559-575.

66 Kheradpour, P. & Kellis, M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucl. Acids Res.*, **42**, 2976-2987.

67 Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D. and van Helden, J. (2011) RSAT 2011: regulatory sequence analysis tools. *Nucl. Acids Res.*, **39**, W86-W91.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

FIGURE LEGENDS

Figure 1: Identification of the *long-range* meQTL with GWAS of PC1 in the SYS adolescents (top) and parents (bottom). Separate analyses were conducted in the samples of SYS adolescents (n=132) and SYS parents (n=278). Pink shaded area indicates the region of CpGs loading into PC1. In both the top and bottom panels, the purple dot indicates the top SNP associated with DNAm (PC1) in the discovery sample of SYS adolescents.

Figure 2: The relative contributions of the *long-range* and *local* meQTLs to methylation of CpGs within the studied genomic region. The contributions are shown as proportions of 'variance explained (r^2) by the *long-range* meQTL (purple lines) and respective *local* meQTLs (yellow lines with stars). Grey triangles indicate the combined contribution of the *long-range* and respective *local* meQTLs at each assessed CpG. The top panel includes the 25 highly variable CpGs loading into PC1 (including 19 polymorphic CpGs). The middle panel includes a total of 456 CpGs (including 85 polymorphic CpGs). The bottom panel includes a total 5,663 CpGs (including 273 polymorphic CpGs) located within the region of high linkage disequilibrium with the *long-range* meQTL (as shown in Figure 1).

Figure 3: Genomic landscape of regions containing the *long-range* meQTL (center), *HLA-DRB1* (left) and *HLA-DQB1* (right). DNase-I hypersensitivity signals, histone modification signals for H3K27ac, H3K4Me1 and H3K4Me3, RNA-seq transcription signals, CTCF-binding positions by Chip-seq, CpG islands, and CpGs assessed with the 450K BeadChip are those observed in the lymphoblastoid GM12878 cells (ENCODE/Duke). Binding of transcription factors is that compiled from 72 cell types in Chip-seq ENCODE V2 (capital-G denotes binding in GM12878 cells).

Figure 4: Associations of the *long-range* meQTL with mRNA expression (top) and hypothesized models (bottom) of how minor and major alleles of the *long-range* meQTL impact regional chromatin structure and mRNA expression. The data are plotted by the meQTL genotype. The *long-range* meQTL is located within a super-enhancer. Depending on its allele, the super-enhancer may interact with either the *HLA-DRB1* promoter (minor allele) or the *HLA-DQB1* promoter (major allele). When interacting with the *HLA-DRB1* promoter (minor

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

allele), it could activate mRNA expression of *HLA-DRB1* and 2 neighboring genes transcribed in the same direction (*HLA-DRB6* and *HLA-DRB5*); note that *HLA-DRB6* is a transcribed pseudogene (30). When interacting with the *HLA-DQB1* promoter (major allele), it could create a chromatin loop and activates mRNA expression of *HLA-DQB1* but not that of the neighboring *HLA-DQA1* because that gene is transcribed in the opposite direction.

Figure 5. Pairwise linkage disequilibrium between the *long-range* meQTL and previously identified disease loci (the National Human Genome Research Institute [NHGRI] GWAS Catalogue, only loci with $r^2>0.2$ are shown). LD was calculated with the 1,000G European reference sample (March 2012).

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

TABLES

Table 1. Principal component analysis of the 25 highly variable CpGs from the studied genomic region - PC1 loading in the 4 studied samples

CpG ID	Gene	Position	SYS adolescents	SYS parents	IMAGEN	OFCCR
cg17369694[a]	*HLA-DRB5*	32,485,396	0.94	0.93	0.93	0.91
cg01341801[a]	*HLA-DRB5*	32,489,203	0.95	0.94	0.95	0.96
cg23365293[a]	*HLA-DRB5*	32,489,984	0.83	0.76	0.74	0.73
cg08265274[a]	*HLA-DRB5*	32,490,444	0.74	Not loading	Not loading	0.46
cg22627029[a]	*HLA-DRB6*	32,520,615	0.82	0.88	0.90	0.87
cg25140213[a]	*HLA-DRB6*	32,522,683	0.94	0.92	0.94	0.93
cg10995422[a]	*HLA-DRB6*	32,522,872	0.59	NA	0.73	0.68
cg24638099[a]	*HLA-DRB6*	32,526,027	0.59	0.53	0.55	0.59
cg11752699[a]	*HLA-DRB6*	32,526,669	0.95	0.94	0.95	0.95
cg26590106[a]	*HLA-DRB1*	32,548,321	0.91	NA	0.90	0.9
cg11404906	*HLA-DRB1*	32,551,749	-0.88	-0.93	-0.94	-0.96
cg09139047[a]	*HLA-DRB1*	32,552,042	-0.93	-0.87	-0.88	-0.93
cg15602423[a]	*HLA-DRB1*	32,552,095	-0.72	-0.68	-0.84	-0.91
cg00211215	*HLA-DRB1*	32,552,246	-0.78	-0.68	-0.70	-0.85
cg09949906[a]	*HLA-DRB1*	32,552,350	-0.68	-0.53	-0.72	-0.87
cg22933800	*HLA-DQA1*	32,605,704	0.60	0.63	0.70	0.66
cg24470466[a]	*HLA-DQA1*	32,608,858	0.56	0.47	0.59	0.57
cg11784298[a]	*HLA-DQA1*	32,610,971	0.53	0.53	0.64	0.5
cg14323910	*HLA-DQB1*	32,628,305	0.52	0.52	0.60	0.53
cg10180404[a]	*HLA-DQB1*	32,632,334	0.67	0.68	0.67	0.67
cg21493951	*HLA-DQB1*	32,632,338	0.60	0.57	0.62	0.58
cg13423887	*HLA-DQB1*	32,632,694	-0.55	-0.61	-0.74	-0.76
cg18572898[a]	*HLA-DQA2*	32,712,103	0.53	0.43	Not loading	Not loading
cg07389699[a]	*HLA-DQB2*	32,728,786	0.49	0.42	Not loading	Not loading
cg24080129[a]	*TAP2*	32,797,488	0.54	Not loading	0.49	0.43

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

[a] Polymorphic CpGs**.** Dark-grey shading indicates CpGs that loaded negatively into PC1 in all four samples, whereas light-grey shading indicates CpGs that loaded positively into PC1 in all four samples.

SYS – The Saguenay Youth Study

IMAGEN – the IMAGEN Study

OFCCR – the OFCCR Study

NA – CpGs not available for the study, as they did not pass quality control

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

Table 2. Replication of the GWAS-identified *long-range* meQTL in IMAGEN and OFCCR samples

Sample	SNP	Minor allele (frequency)	Estimate (SE)	P
SYS adolescents (n=132)	rs4959030	A (0.13)	6.8 (0.8)	$2.8x10^{-17}$
SYS parents (n=278)	rs4959030	A (0.16)	5.7 (0.5)	$3.1x10^{-31}$
IMAGEN (n=639)	rs9271366	G (0.14)	5.3 (0.2)	$4.0x10^{-71}$
OFCCR (n=1,747)	rs9270986	A (0.15)	5.5 (0.1)	$5.2\ x10^{-199}$

As the genotypes of the SYS-identified *long-range* meQTL (rs4959030) were not available in either IMAGEN or OFCCR, the replication analyses were carried out with SNPs that were in closest linkage disequilibrium (LD) with the index SNP rs4959030; these were rs9271366 (r^2=0.76) and rs9270986 (r^2=0.68) for the IMAGEN and OFCCR, respectively. LD was determined based on the 1,000G data from March 2012.

Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

Table 3. IMAGEN – associations of the *long-range* meQTL[a] with gene expression

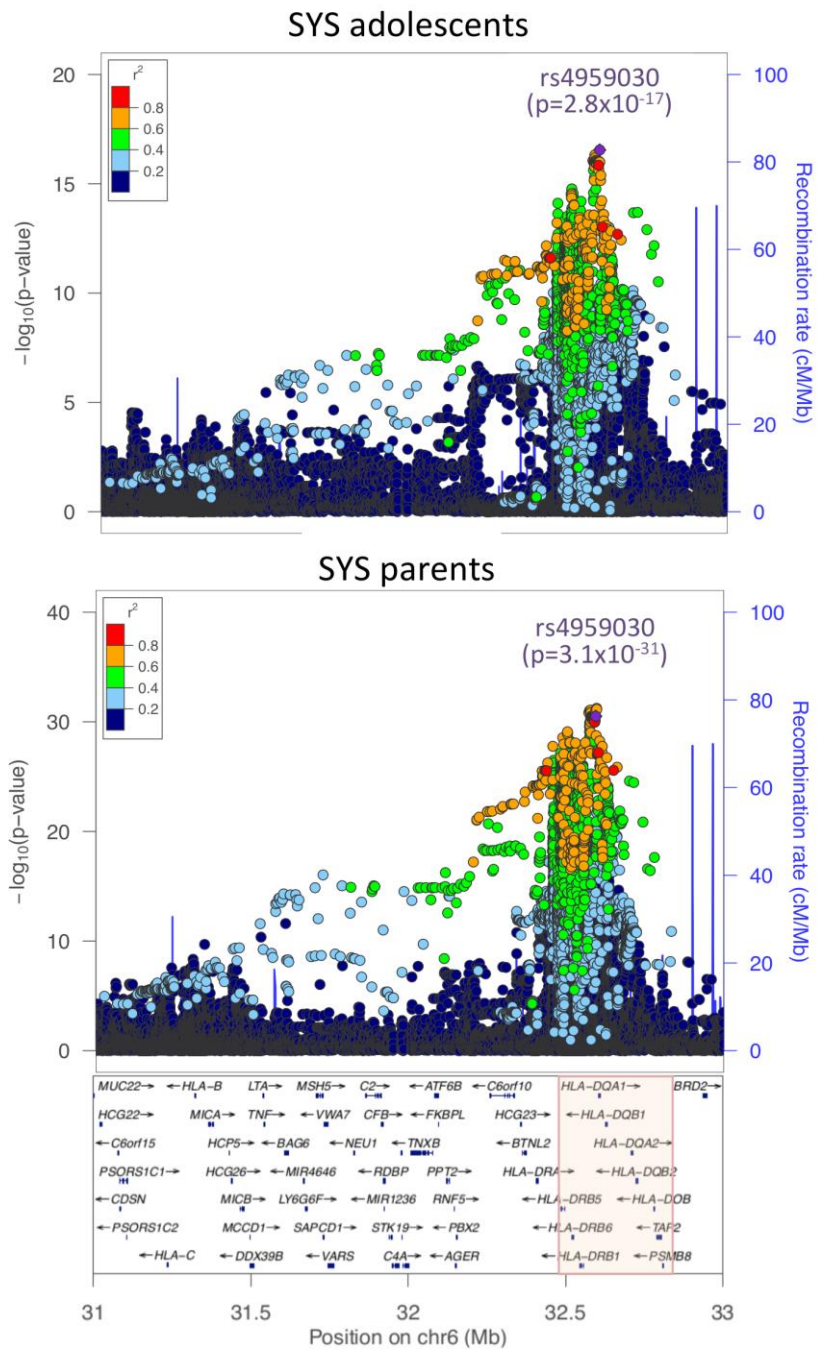
Gene			Causal inference test[c]
	Estimate (SE[b])	P	P
HLA-DRB5	1174.4 (39.7)	1.0E-123	7.1E-20
HLA-DRB1	660.0 (34.1)	3.5E-48	2.4E-13
HLA-DQA1	27.5 (20.6)	2.4E-01	1.0E+00
HLA-DQB1	-64.7 (6.3)	7.1E-18	7.5E-17

[a] As rs4959030 (the GWAS-identified meQTL in the SYS sample) was not available in IMAGEN, the DNAm and mRNA expression analyses in IMAGEN were carried out with rs927136, a SNP in closest linkage disequilibrium (r^2=0.76) with rs4959030.

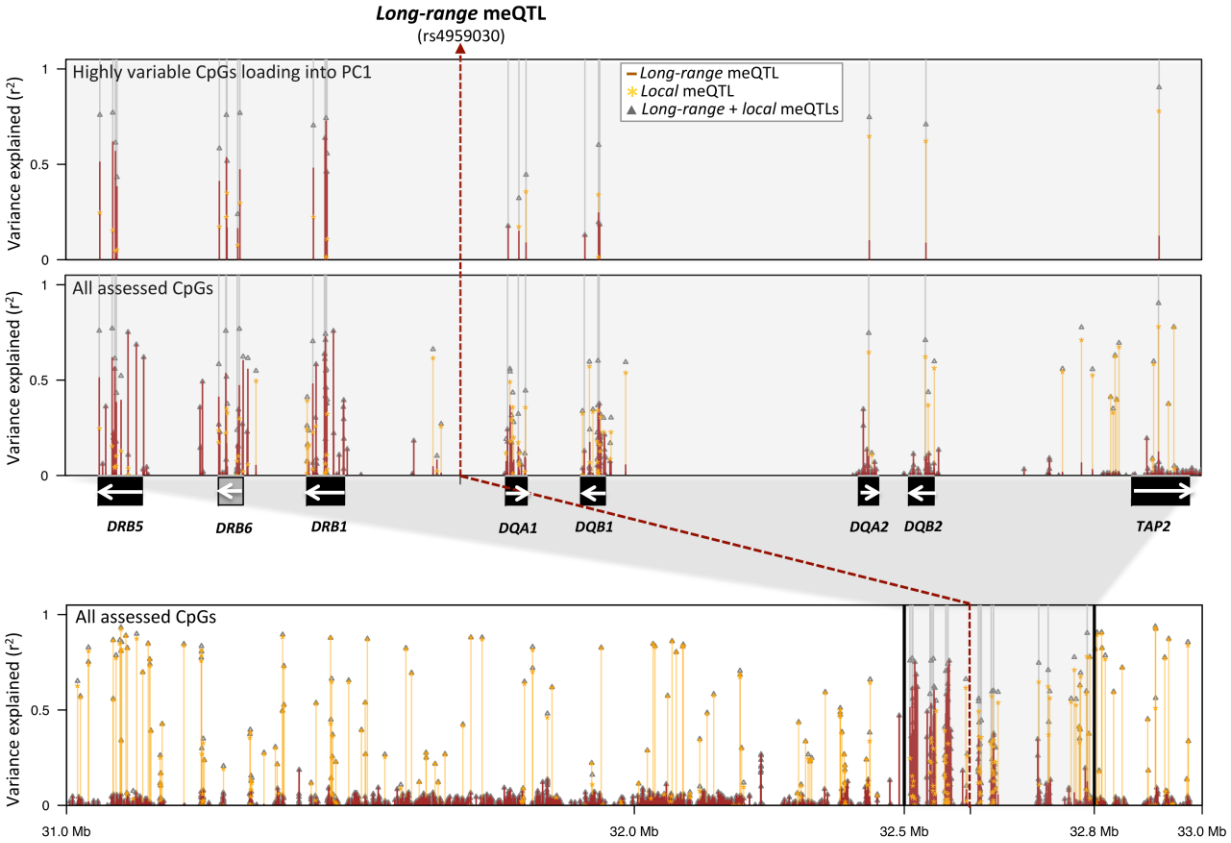
[b] Bootstrap standard errors based on 5,000 replications.

[c] Causal inference test (35).

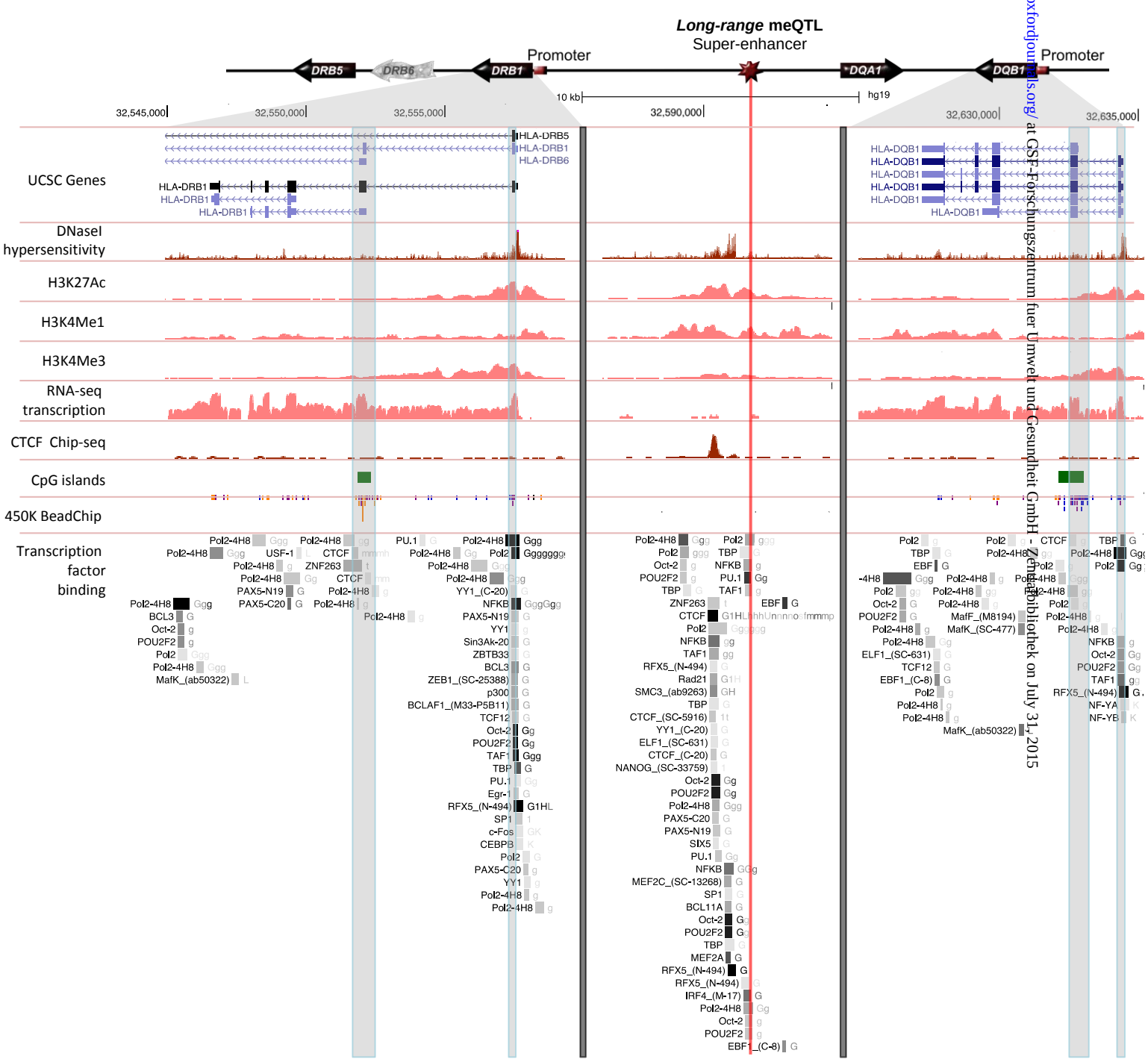
Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015



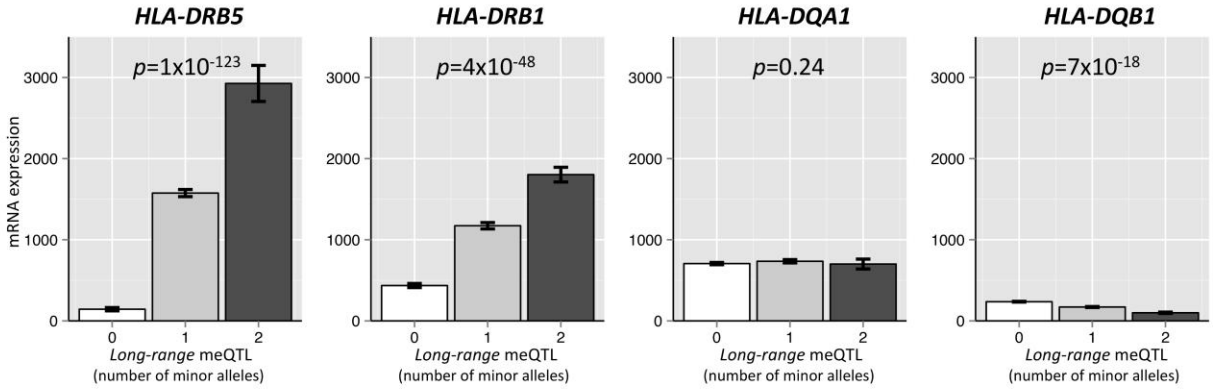
Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

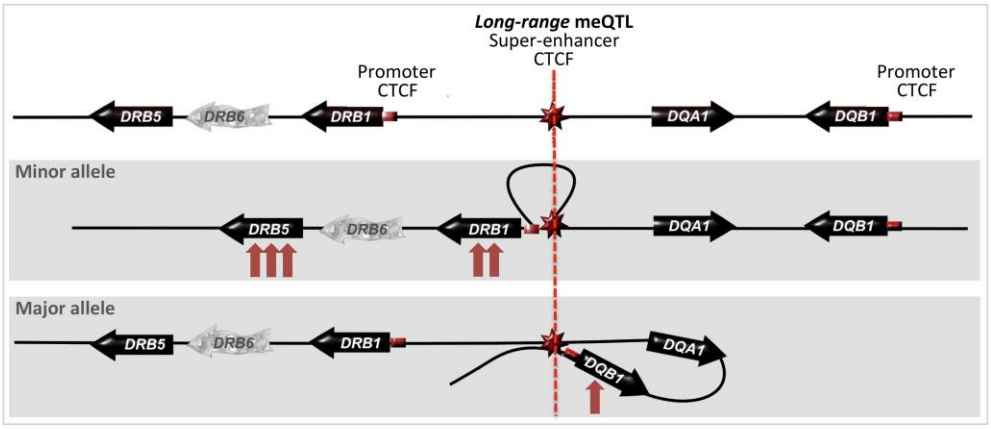


Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015

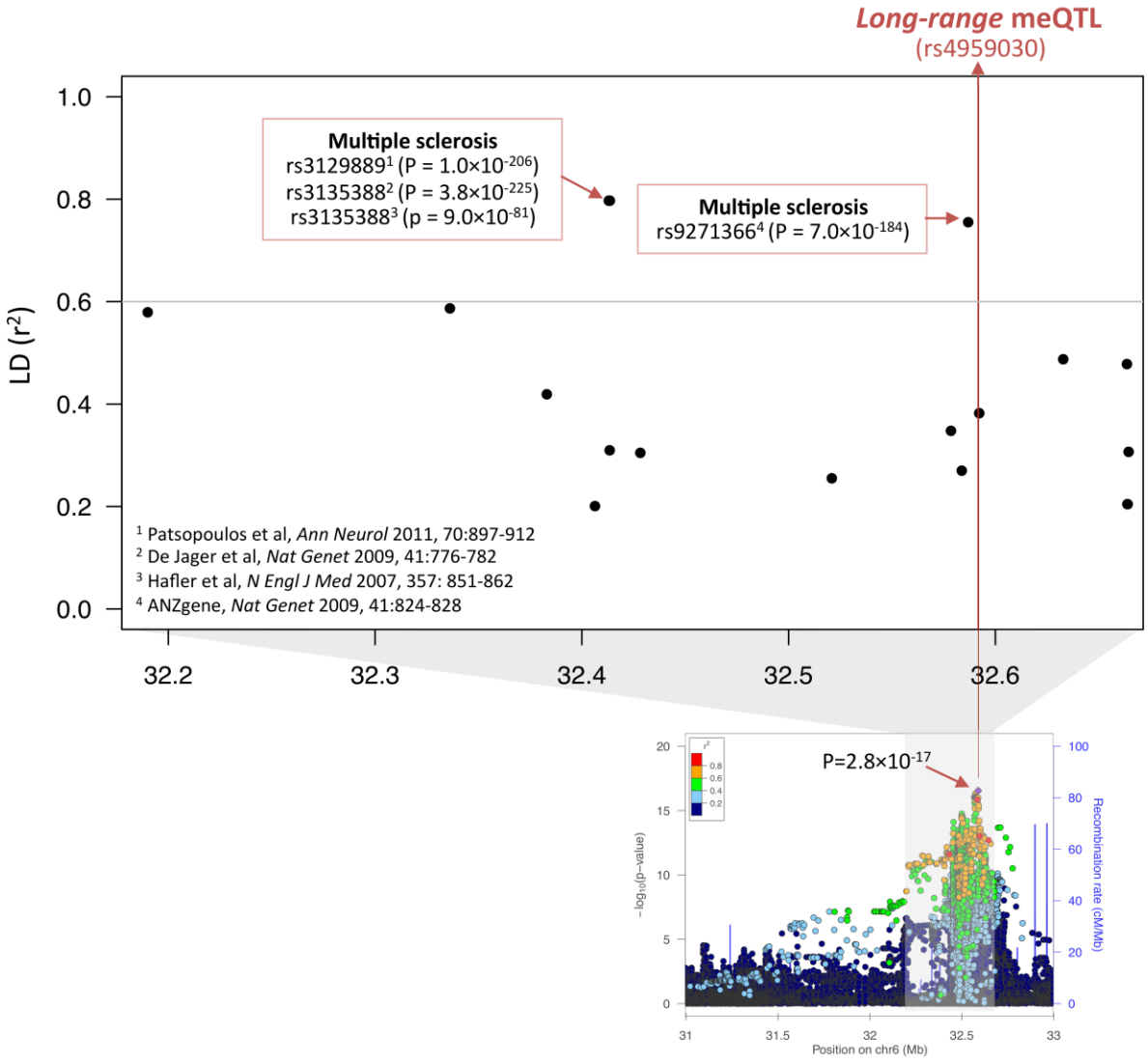


http://epigenomesmb.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015





Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015



Downloaded from http://hmg.oxfordjournals.org/ at GSF-Forschungszentrum fuer Umwelt und Gesundheit GmbH - Zentralbibliothek on July 31, 2015