



## A pooling-based approach to mapping genetic variants associated with DNA methylation

Irene M. Kaplow, Julia L. Maclsaac, Sarah M. Mah, et al.

*Genome Res.* 2015 25: 907-917 originally published online April 24, 2015

Access the most recent version at doi:[10.1101/gr.183749.114](https://doi.org/10.1101/gr.183749.114)

---

**References** This article cites 88 articles, 15 of which can be accessed free at:  
<http://genome.cshlp.org/content/25/6/907.full.html#ref-list-1>

**Creative Commons License** This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

© 2015 Kaplow et al.; Published by Cold Spring Harbor Laboratory Press

## Method

# A pooling-based approach to mapping genetic variants associated with DNA methylation

Irene M. Kaplow,<sup>1,2</sup> Julia L. Maclsaac,<sup>3,5</sup> Sarah M. Mah,<sup>3,5</sup> Lisa M. McEwen,<sup>3,4</sup> Michael S. Kobor,<sup>3,4</sup> and Hunter B. Fraser<sup>2</sup>

<sup>1</sup>Department of Computer Science, Stanford University, Stanford, California 94305, USA; <sup>2</sup>Department of Biology, Stanford University, Stanford, California 94305, USA; <sup>3</sup>Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, University of British Columbia, Vancouver, British Columbia V5Z 4H4, Canada; <sup>4</sup>Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia V5Z 4H4, Canada

DNA methylation is an epigenetic modification that plays a key role in gene regulation. Previous studies have investigated its genetic basis by mapping genetic variants that are associated with DNA methylation at specific sites, but these have been limited to microarrays that cover <2% of the genome and cannot account for allele-specific methylation (ASM). Other studies have performed whole-genome bisulfite sequencing on a few individuals, but these lack statistical power to identify variants associated with DNA methylation. We present a novel approach in which bisulfite-treated DNA from many individuals is sequenced together in a single pool, resulting in a truly genome-wide map of DNA methylation. Compared to methods that do not account for ASM, our approach increases statistical power to detect associations while sharply reducing cost, effort, and experimental variability. As a proof of concept, we generated deep sequencing data from a pool of 60 human cell lines; we evaluated almost twice as many CpGs as the largest microarray studies and identified more than 2000 genetic variants associated with DNA methylation. We found that these variants are highly enriched for associations with chromatin accessibility and CTCF binding but are less likely to be associated with traits indirectly linked to DNA, such as gene expression and disease phenotypes. In summary, our approach allows genome-wide mapping of genetic variants associated with DNA methylation in any tissue of any species, without the need for individual-level genotype or methylation data.

[Supplemental material is available for this article.]

DNA methylation is an epigenetic mark that usually occurs at cytosine bases within CG dinucleotides (CpGs) in the human genome. CpGs often occur in dense clusters, known as CpG islands, which are surrounded by regions known as CpG shores; in nonmalignant cells, CpG islands and shores are less frequently methylated than CpGs outside of these regions (Smith and Meissner 2013). Methylated CpG islands and shores in promoters may provide a “locking mechanism” that prevents repressed genes from being reactivated, while methylation in gene bodies is often associated with active transcription (Wolf et al. 1984; Jones 2012). In addition, DNA methylation has been linked to a wide range of diseases, including cancer, Alzheimer’s disease, bipolar disorder, and type 2 diabetes (Baylin et al. 1998; Baylin and Herman 2000; Gamazon et al. 2012; Irier and Jin 2012; The Cancer Genome Atlas Network 2012; Dayeh et al. 2013; Pease et al. 2013; Ambrosone et al. 2014; De Jager et al. 2014; Lunnon et al. 2014). Interestingly, CpGs whose methylation has been associated with gene expression and disease are found not only in promoter regions or gene bodies but also in other parts of the genome, such as enhancers and insulators, suggesting additional roles of DNA methylation in transcriptional regulation (You et al. 2011; Jones 2012; Gutierrez-Arcelus et al. 2013; Banovich et al. 2014; Zhang et al. 2014). Although many studies have investigated potential additional roles, general conclusions

about the role of methylation outside of promoters and gene bodies are still lacking (Jones 2012).

Several studies have investigated the relationship between DNA methylation and other epigenetic factors, such as histone modifications and chromatin accessibility (Wrzodek et al. 2012; Shi et al. 2014; Wagner et al. 2014; Zhang et al. 2014). For example, DNA methylation is associated with transcription factor (TF) binding (Thomson et al. 2010; Wiench et al. 2011; You et al. 2011; Feldmann et al. 2013; Ziller et al. 2013; Heyn 2014; Shi et al. 2014; Smith et al. 2014). In associations with epigenetic variation, the direction of causality is usually unclear; DNA methylation may affect TF binding or may be affected by it, or both may be determined by another factor (or any combination of these).

In addition to epigenetic variation, DNA methylation can also be associated with genetic variation (Gibbs et al. 2010; Bell et al. 2011, 2012; Bibikova et al. 2011; Fraser et al. 2012; Lam et al. 2012; Drong et al. 2013; Grundberg et al. 2013; Gutierrez-Arcelus et al. 2013; Heyn et al. 2013; Liu et al. 2013; Moen et al. 2013; Zhi et al. 2013; Ambrosone et al. 2014; Banovich et al. 2014; De Jager et al. 2014; Lunnon et al. 2014; Shi et al. 2014; Smith et al. 2014; Wagner et al. 2014; Zhang et al. 2014). Associations with genetic variants such as SNPs are qualitatively different from epigenetic associations with disease or gene

<sup>5</sup>These authors contributed equally to this work.

Corresponding author: [hbfraser@stanford.edu](mailto:hbfraser@stanford.edu)

Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.183749.114>.

© 2015 Kaplow et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

expression because the causality is clear: Mendelian randomization ensures that an individual's genotype is a random combination of parental alleles, and thus any associations must be due to the effects of genotype (in a study free of confounding factors) (Mokry et al. 2014). Genetic variants showing these associations are known as DNA methylation quantitative trait loci (mQTLs).

Nearly all published human mQTL studies have relied upon commercially available microarrays that interrogate either approximately 28,000 CpGs or approximately 480,000 CpGs, corresponding to 0.1% and 1.7% of CpGs in the genome, respectively (Gibbs et al. 2010; Bell et al. 2011, 2012; Fraser et al. 2012; Drong et al. 2013; Grundberg et al. 2013; Gutierrez-Arcelus et al. 2013; Heyn et al. 2013; Liu et al. 2013; Moen et al. 2013; Zhi et al. 2013; Ambrosone et al. 2014; Banovich et al. 2014; Shi et al. 2014; Smith et al. 2014; Wagner et al. 2014; Zhang et al. 2014). Not surprisingly, the number of CpGs associated with mQTLs is generally greater in the studies using the larger arrays. However, the remaining 98.3% of CpG sites, including nearly all CpGs outside of CpG islands and shores, have yet to be included in any mQTL study, thus limiting our understanding of the roles of DNA methylation far from CpG islands.

In addition to covering only a small subset of the genome, DNA methylation arrays can be affected by experimental variability between samples, such as differences between the quality of bisulfite treatment across samples and “batch effects” that affect groups of samples, such as the day of hybridization (Leek et al. 2010; Sun et al. 2011). Although many computational methods have been developed to mitigate these issues, they can be impossible to correct for perfectly and thus remain important potential sources of both false positives and false negatives (Johnson et al. 2007; Leek et al. 2010; Sun et al. 2011; Stegle et al. 2012; Mostafavi et al. 2013; Yousefi et al. 2013).

Another property of DNA methylation arrays is that they provide only the average methylation level for each targeted CpG in a sample. If an individual's two alleles have the same methylation level, this is not an issue; however, samples heterozygous for a *cis*-acting mQTL will have allele-specific methylation (ASM) because *cis*-acting variants affect methylation of only the CpG allele to which they are linked (Kerkel et al. 2008; Shoemaker et al. 2010). Thus, a significant source of information present in heterozygotes is lost. A previous ChIP-seq study used allele-specific information with a combined haplotype test, in which the authors modeled the read depth for each allele from each individual and tested whether there is a significant difference between allelic read depths across individuals, identifying thousands of histone modification QTLs (McVicker et al. 2013). Other ChIP-seq studies have also identified allele-specific histone modifications (Mikkelsen et al. 2007; Kasowski et al. 2013; Kilpinen et al. 2013), underscoring the value of allele-specific information.

Deep sequencing of bisulfite-converted DNA (Frommer et al. 1992) overcomes many of the limitations of microarrays by allowing genome-wide detection of DNA methylation and ASM (Lister et al. 2009, 2013; Xie et al. 2012; Schmitz et al. 2013a,b). Applying this technology to F1 hybrid mice (CAST $\times$ 129) revealed more than 10<sup>5</sup> sites of allele-specific DNA methylation, suggesting that *cis*-acting effects of genetic variation on DNA methylation are widespread (Xie et al. 2012). However, performing whole-genome bisulfite sequencing—or even reduced representation bisulfite sequencing, which preferentially targets CpG-rich regions (Meissner et al. 2005, 2008)—on more than a few human individuals is prohibitively expensive. Nevertheless, because distal noncoding variants affecting DNA methylation may play key roles

in some human traits (Ward and Kellis 2012; Dayeh et al. 2013; Grundberg et al. 2013), whole-genome bisulfite sequencing of many individuals may prove to be essential in understanding the genetic and epigenetic bases of phenotypic variation.

In this study, we developed a novel method for identifying mQTLs from pooled sequencing data that allows genome-wide mQTL mapping while reducing the effort, cost, and experimental variability associated with bisulfite sequencing and genotyping of many individual samples. Our approach allows us to study thousands of molecular traits in a single experiment, in contrast to existing pooled QTL mapping methods that are limited to individual phenotypes (Michelmore and Paran 1991; Ehrenreich et al. 2010). We first tested our approach in simulations and then applied it to bisulfite sequencing data from a pool of 60 human lymphoblastoid cell lines (LCLs). We identified more than 2000 novel mQTLs, including some that are also associated with variation in DNase hypersensitivity, TF binding, gene expression levels, or complex diseases (Pickrell et al. 2010; Degner et al. 2012; Lappalainen et al. 2013; Ding et al. 2014; Welter et al. 2014). We also show that TF binding sites and open chromatin regions from LCLs are enriched for mQTLs. Our approach represents a powerful and cost-effective framework for mapping mQTLs genome wide, in any species.

## Results

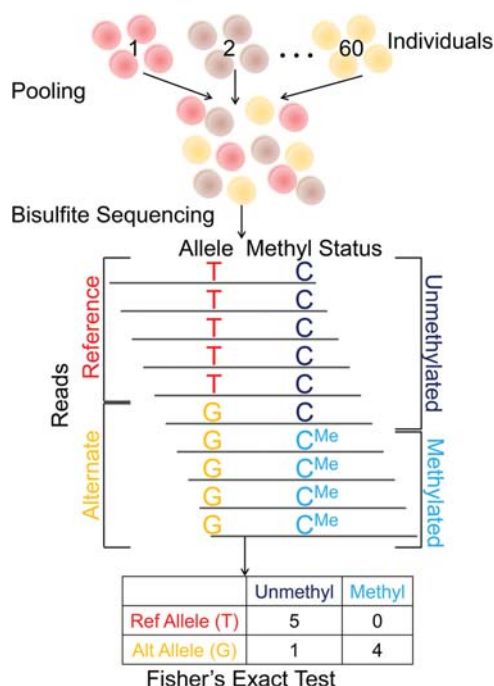
### A pooling approach to mQTL mapping

In most genetic association studies of quantitative traits, each sample is genotyped at many variants (typically at least 10<sup>6</sup> for humans), and trait values (e.g., DNA methylation at specific CpGs) are measured for each sample. Genotypes are then compared to the trait values to detect statistically significant associations. This approach has been widely used in many species to map loci associated with both molecular-level traits (expression QTLs, mQTLs, DNase hypersensitivity QTLs, etc.) and organismal-level traits (height, blood pressure, etc.).

Our pooling approach is outlined in Figure 1. The central idea is that, for any *cis*-acting mQTL-CpG pair, if both the variant and the CpG are observed on the same DNA sequencing read, then any association between the two can be detected using Fisher's exact test (see Methods). For example, a CpG near a G/T SNP might have a higher methylation level when linked to the G allele than when linked to the T allele. Analysis at the level of alleles, rather than individuals, allows samples to be pooled prior to bisulfite treatment because the allelic identity of every informative read can be inferred directly from the read's sequence. In addition to minimizing experimental variability between samples, pooling obviates the need for individual-level genotyping and DNA methylation profiling.

### Assessing the pooling approach via simulations

We compared our pooling approach (“pooled ASM method”), which allows us to detect allele-specific methylation (ASM), to a more traditional approach (“traditional non-ASM method”), which consists of bisulfite sequencing and genotyping of each sample separately, followed by the comparison of the individual-level genotypes and average DNA methylation levels. We focused on genome-wide methods and therefore did not include DNA methylation arrays in the simulations. Each simulation featured an mQTL-CpG pair, for which we estimated the power to detect the mQTL (at  $P < 0.001$ ). Across simulations we varied five



**Figure 1.** We use pooled bisulfite sequencing to identify mQTLs. Our pooling method enables us to identify mQTLs directly from bisulfite sequencing reads. In this approach, cells or DNA samples from all individuals are combined into a single pool, which is then subject to bisulfite sequencing. Alleles and methylation statuses are inferred from the sequence reads, which are then used to generate a  $2 \times 2$  contingency table (where columns represent methylation statuses and rows represent alleles). Fisher's exact test is used to compute a  $P$ -value for the null hypothesis of no association.

parameters: the effect size of the mQTL (the correlation between allele and methylation status); the read depth at the site (across all individuals); the minor allele frequency (MAF); the minor DNA methylation status frequency (the frequency of the less common methylation status); and the number of individuals. We also ran our simulations at additional  $P$ -value cutoffs and simulated variant-CpG pairs with no association between allele and methylation status at various  $P$ -value cutoffs to evaluate our false positive rate.

The pooled ASM method identified substantially more mQTLs than the traditional non-ASM method at most parameter values—the exception being a strong mQTL covered by many reads, in which case the methods both had close to 100% power (Fig. 2; Supplemental Figs. 1–8). In fact, for effect sizes  $\leq 0.83$ , read depths  $\leq 40\times$ , and  $\leq 100$  individuals, the pooled ASM method had over twice the power of the traditional non-ASM method. The false positive rate for both methods was extremely low, especially for lower  $P$ -value cutoffs (Supplemental Figs. 9–12). In addition, for nearly all parameter settings, the pooled ASM method produced significantly lower mQTL  $P$ -values than the traditional method (Supplemental Fig. 13). We found similar results using two alternative statistical approaches to analyze the simulated data (Supplemental Figs. 3, 4).

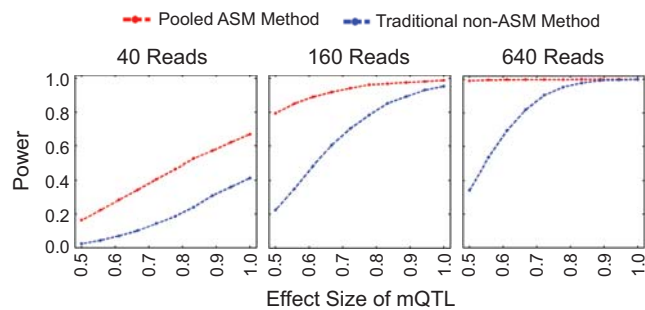
The increased power of our approach is due primarily to the difference in what information is extracted from heterozygotes. For example, consider a strong mQTL (0% methylation of the major allele, 100% for the minor allele) with a low MAF that is present as only either major-allele homozygotes or heterozygotes in a particular cohort. The traditional non-ASM method averages together

the effects of the two different alleles in each heterozygote, thus diluting the signal from each allele (the traditional method cannot distinguish between when the minor allele in heterozygotes always corresponds to the minor methylation status and when the allele in heterozygotes has no correlation with the methylation status). As a result, the two quantities being compared—average methylation of the CpG in homozygotes versus heterozygotes—would be 0% versus 50%. In contrast, because the pooled ASM method takes into account the allelic origin of each read, it is equally informative regardless of whether a read is from a homozygote or heterozygote. In our example, the two quantities in the comparison—methylation of the CpG in one allele versus the other—would be 0% versus 100% in the pooled ASM method, which is more easily detected with limited data. This concept applies to any mQTL that can be found in a heterozygous state, even when all three possible genotypes are present in a cohort. As a result, the pooled ASM method had the greatest advantage at large MAFs (Supplemental Figs. 5, 6, 11, 12) because these have the most heterozygotes.

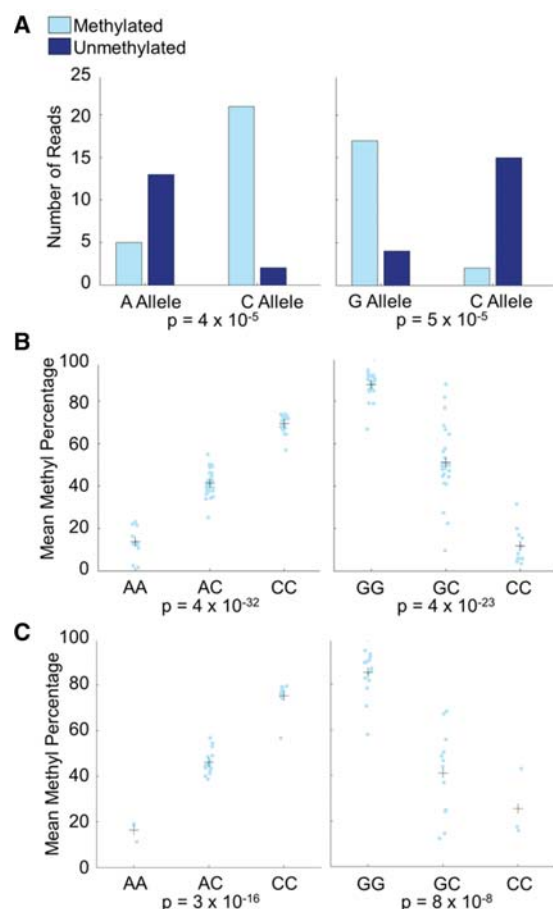
### Applying our pooling approach to empirical data

As an experimental test of our pooling approach, we performed bisulfite sequencing on pooled genomic DNA from LCLs derived from 60 Yoruban (YRI) HapMap individuals (see Methods) (Frazer et al. 2007). We obtained  $\sim 860$  million pairs of 101-bp reads, of which 77.1% passed our quality control filters and mapped uniquely, corresponding to an average per-base coverage of  $\sim 40\times$ .

To assess the efficiency of our bisulfite treatment and pooling, we performed three quality control tests. First, to evaluate the efficiency of bisulfite conversion of unmethylated cytosines to thymines, we spiked unmethylated lambda phage DNA into the pool prior to bisulfite treatment. The percentage of unconverted lambda phage cytosines ranged from 0.0% to 0.2% across library/sequencing lane pairs, suggesting that the bisulfite treatment consistently had at least 99.8% efficiency, which is comparable to previous bisulfite sequencing studies (Lister et al. 2009; Banovich et al. 2014). Second, we estimated the rate of spurious cytosine to thymine conversion by creating two untreated



**Figure 2.** Simulations comparing our pooled ASM method to the traditional non-ASM method. Our pooled ASM method has more power to detect mQTLs than the traditional non-ASM method, especially for low effect sizes. Simulated reads from 100 individuals were sampled from a negative binomial distribution, with minor allele and minor methylation frequencies of 0.1. Power to detect an mQTL is shown for each method as a function of mQTL effect size and total read depth. The effect size is the correlation between allele and methylation status, and the power is the fraction of the simulations in which we identified an mQTL (with  $P < 0.001$ ). Additional simulations are shown in Supplemental Figures 1–8, and ROC curves from the simulations are shown in Supplemental Figures 9–12.



**Figure 3.** Pyrosequencing validation of mQTLs. Shown are two mQTLs involving SNPs previously identified in GWAS. (*Left*) The SNP is associated with age-related macular degeneration; (*right*) the SNP is associated with the ratio of visceral adipose tissue to subcutaneous adipose tissue. (A) Pooled bisulfite sequencing for the two mQTLs, showing strong association. (B) Pyrosequencing validation of the two mQTLs in individual samples that were used for the pooled bisulfite sequencing confirms the bisulfite sequencing results. Light blue points are the methylation percentages from individuals, and crosses are the mean methylation percentages for individuals of each genotype. (C) Pyrosequencing validation of the two mQTLs in 30 additional YRI individuals shows that the mQTLs are not limited to the individuals in our study. Light blue points are the methylation percentages from individuals, and crosses are the mean methylation percentages for individuals of each genotype.

control libraries from our pooled human samples. In these libraries, only 0.4%–0.5% of bases that mapped to cytosines were thymines, suggesting a low rate of spurious C-to-T conversion, sequencing errors, and unannotated C/T SNPs. Finally, we estimated the relative abundance of each individual sample's DNA in the pool and found that it was near the expected 1:60 ratio (1.0%–2.8% abundance of each sample), confirming that our pool was relatively homogenous (Supplemental Fig. 14). We note that sample heterogeneity should not impact our ability to map mQTLs so long as each allele is well represented, because our approach does not require an equal abundance of each sample in the pool.

Our pooled data provided sufficient coverage to test the strength of association at 823,726 variant-CpG pairs, from which we identified 2379 mQTLs (at 2332 unique CpGs) at  $P < 0.001$  (corresponding to a false discovery rate [FDR] of 6.0%) (see Methods; Supplemental Table 6). Examples of two mQTLs are shown in Figure 3A. mQTLs tended to be close to their corresponding CpGs because our method required that they be on the same sequenced fragment; the average distance was 25.4 bp, with a range of 0–377 bp (Supplemental Fig. 15).

We used pyrosequencing on individual samples to validate our results in two ways. First, as a technical validation of the pooling method, we performed pyrosequencing at two CpGs with mQTLs in the same 60 samples that composed our pool. Both of these were successfully validated in the unpooled samples (Fig. 3B). Second, we tested the reproducibility of our mQTLs by pyrosequencing eight CpGs (spanning a wide range of  $P$ -values and including the two with technical validation) in a separate set of 30 YRI samples, which are offspring of the 60 individuals in our pool. In seven of eight cases, the mQTL was successfully validated in these additional samples (Fig. 3C; Table 1; Supplemental Figs. 16–21; also described below).

We then compared our mQTLs to those in two recent studies of YRI LCLs (Banovich et al. 2014; Zhang et al. 2014), which both used the largest commercially available DNA methylation array, the Illumina Infinium HumanMethylation450 BeadChip (Bibikova et al. 2011). Comparing these studies to one another, we found 19.1%–45.7% overlap (depending on the direction of analysis; see Methods). The disagreement was likely caused by a combination of false positives and false negatives (e.g., due to low power or between-sample variability). Comparing these two catalogs to our data, we found that <1% of our approximately 800K tested CpG sites were also tested in each of these studies.

**Table 1.** Pyrosequencing validation of mQTLs

mQTL rsID	CpG position	mQTL annotation	$P$ -value in pooled bisulfite sequencing data	$P$ -value in pyrosequencing data
rs10737680	Chr 1:196679457	GWAS hit	$4.3 \times 10^{-5}$	$3.6 \times 10^{-32}$ , $3.3 \times 10^{-16}$
rs9517668	Chr 13:99923789	In LD with GWAS hit	$7.7 \times 10^{-4}$	>1
rs12905925	Chr 15:65164853	dsQTL	$5.3 \times 10^{-6}$	$4.0 \times 10^{-7}$
rs617201	Chr 17:3089145	In LD with GWAS hit	$3.5 \times 10^{-5}$	$6.4 \times 10^{-9}$
rs1113144	Chr 18:71774787	In LD with GWAS hit	$1.4 \times 10^{-5}$	$4.4 \times 10^{-13}$
rs4809456	Chr 20:61660870	In LD with GWAS hit	$1.2 \times 10^{-4}$	$2.4 \times 10^{-5}$
rs62223713	Chr 21:30158046	Exon-level eQTL	$2.0 \times 10^{-4}$	$9.7 \times 10^{-12}$
rs7705033	Chr 5:122774795	GWAS hit	$4.7 \times 10^{-5}$	$3.8 \times 10^{-23}$ , $8.3 \times 10^{-8}$

Seven of eight tested mQTLs were successfully validated using pyrosequencing. All mQTLs have  $P$ -values for the pyrosequencing validation on 30 YRI individuals who were not included in our pooled bisulfite sequencing data. mQTLs with two  $P$ -values are those that were also validated using the 60 YRI individuals in our study; the  $P$ -value from pyrosequencing of the 60 individuals in our data is listed first. All pyrosequencing data  $P$ -values have been Bonferroni corrected.

Focusing on the small number of sites in common, we found a similar level of overlap (40.0%–53.3%; 4/10 and 8/15, respectively), which was significantly more than expected by chance (hypergeometric  $P=5.5 \times 10^{-3}$  and  $9.6 \times 10^{-7}$ , respectively). In sum, our pooling-based approach agrees with microarray studies as well as these studies agree with one another, despite the major methodological differences between the approaches (Supplemental Figs. 22, 23).

### mQTLs were associated with molecular-level and organismal-level traits

To explore the potential effects of our mQTLs, we examined their distribution in CpG islands and shores as well as in genomic regions with different chromatin states (see Methods). Unlike DNA methylation microarrays that are primarily targeted in and around CpG islands, only 2.0% of our tested CpGs were in islands, and only 10.8% were within 2 kb of CpG islands (Irizarry et al. 2009; Wu et al. 2010). Instead, most (86.0%) were in repressed/inactive genomic regions, which usually lack CpG islands. As a result, the majority (78.1%) of the CpGs with mQTLs was also in repressed/inactive regions, although this is less than the 86% expected by chance, suggesting that variants in active genomic regions may be more likely to influence methylation than those in inactive regions. Chromatin states that are related to active transcription and active enhancers were enriched for CpGs with mQTLs (Table 2; Supplemental Fig. 24), suggesting that some of these mQTLs may affect transcription or enhancer activity.

To understand if our mQTLs may be involved in cell-type-specific open chromatin, we compared our mQTL catalog with 402 DNase hypersensitivity experiments from the ENCODE and Epigenomics Roadmap projects, which measured chromatin accessibility (Bernstein et al. 2010; The ENCODE Project Consortium 2012; Thurman et al. 2012). We found that our mQTLs were more strongly enriched in regions of open chromatin in LCLs than in open chromatin in other cell types ( $P=1.4 \times 10^{-7}$ ). A total of 9.9% of mQTLs are in open chromatin sites in LCLs, but only

4.9% of tested variants are in open chromatin sites from LCLs (fold enrichment = 2.0); for all other cell types, 48.8% of mQTLs are found in open chromatin regions, and 46.7% of tested variants are found in open chromatin regions (fold enrichment = 1.0) (Supplemental Fig. 25). This suggests that variants may be more likely to affect DNA methylation in tissues where they are in accessible chromatin.

To further understand the relationship between transcriptional activity and mQTLs, we intersected our mQTLs with TF binding and histone modification ChIP-seq peaks from GM12878, an LCL studied extensively by the ENCODE Project Consortium (The ENCODE Project Consortium 2012; Gerstein et al. 2012). We found that mQTLs were enriched in the binding sites for 12 TFs in GM12878; 4.6% of mQTLs are in binding sites of at least one of these TFs, whereas only 1.6% of tested variants are in these binding sites (total fold enrichment across 12 TFs = 2.9) (Supplemental Table 1; Supplemental Fig. 26). The strongest of these enrichments is for binding sites of CTCF, which often acts as an insulator by blocking the spread of chromatin states. SNPs associated with CTCF binding were recently mapped in LCLs (Ding et al. 2014); 18 of these are also mQTLs ( $P=2.0 \times 10^{-3}$ ), and increased binding is usually associated with decreased methylation (binomial  $P=7.7 \times 10^{-4}$ ), suggesting that variants affecting CTCF binding can also impact DNA methylation, consistent with their known relationship (Feldmann et al. 2013).

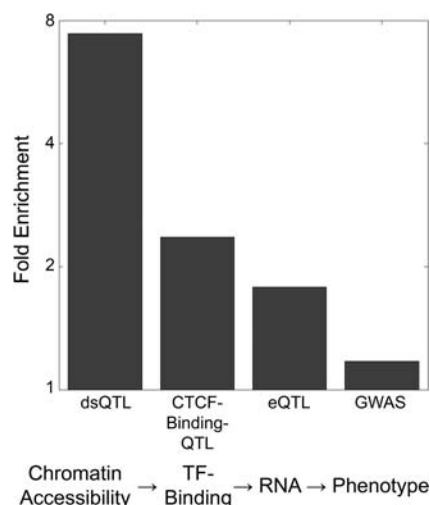
We next investigated the relationship between our mQTLs and other molecular traits by comparing our results to expression QTLs (eQTLs) and DNase hypersensitivity QTLs (dsQTLs) mapped in the same YRI population (Degner et al. 2012; Lappalainen et al. 2013). We found a strong overlap with dsQTLs, 48 of which were also mQTLs ( $P=1.2 \times 10^{-27}$ , fold enrichment = 7.4); alleles associated with lower DNA methylation tended to be associated with open chromatin (binomial  $P=5.5 \times 10^{-7}$ ). For eQTLs, we found that 28 mQTLs are exon-level eQTLs and five are gene-level eQTLs; for exon-level eQTLs, this was significantly more than expected by chance ( $P=1.5 \times 10^{-3}$ , fold enrichment = 1.8). In addition, considering SNPs in strong LD with eQTLs and dsQTLs revealed additional overlaps (Supplemental Tables 2, 3). We validated one eQTL overlap and one dsQTL overlap using pyrosequencing in an additional 30 YRI individuals (Table 1; Supplemental Figs. 17, 21). These overlaps were consistent with our expectation of a relationship between DNA methylation, transcription, and chromatin structure (Gutierrez-Arcelus et al. 2013; Jones et al. 2013).

We also tested to what extent our mQTLs may impact organismal-level traits. To investigate this, we performed a similar comparison of our mQTLs with variants implicated by genome-wide association studies (GWAS) for diseases and other traits (Welter et al. 2014). Nine mQTLs were previously identified in GWAS or are in perfect LD with GWAS hits, and an additional 13 mQTLs are in strong ( $r^2 \geq 0.8$ ) LD with GWAS-implicated variants (Supplemental Table 4; Welter et al. 2014), although this is not significantly more overlaps than expected by chance ( $P=0.67$  for GWAS hits and variants in perfect LD;  $P=0.70$  for GWAS hits and variants in strong LD). We validated two of the overlaps using pyrosequencing in the individuals in our bisulfite sequencing study (Fig. 3B; Table 1) and six of the overlaps in 30 additional YRI individuals; all but one test was successful (Fig. 3C; Table 1; Supplemental Figs. 16, 18–20). All of the validated mQTLs are associated with CpGs that are not included on commercially available microarrays, demonstrating that our method can identify novel disease-associated mQTLs.

**Table 2.** Enrichment of CpGs with mQTLs in LCL chromatin states

Chromatin state	CpGs with mQTLs in state (%)	Enrichment $P$ -value
<b>1. Active TSS</b>	0.90	<b><math>3.0 \times 10^{-6}</math></b>
<b>2. Flanking active TSS</b>	1.0	<b><math>7.4 \times 10^{-3}</math></b>
<b>3. Strong transcription</b>	3.8	<b><math>1.6 \times 10^{-2}</math></b>
<b>4. Weak transcription</b>	3.3	<b><math>1.2 \times 10^{-4}</math></b>
<b>5. Acetylated active enhancer</b>	1.5	<b><math>1.7 \times 10^{-3}</math></b>
6. Acetylated active enhancer (genic)	0.26	>1
<b>7. Active enhancer</b>	2.8	<b><math>4.4 \times 10^{-5}</math></b>
8. Active enhancer (genic)	0.86	$7.8 \times 10^{-2}$
9. Weak enhancer	3.1	$4.9 \times 10^{-1}$
10. Weak enhancer (genic)	0.99	$6.6 \times 10^{-1}$
11. Bivalent TSS	0.30	>1
12. Bivalent enhancer	1.1	>1
13. Repressed Polycomb	23	>1
14. CTCF only	1.9	$7.6 \times 10^{-2}$
15. Quiescent/low	55	>1

Most CpGs with mQTLs occur in regions of the genome that are in the “Repressed Polycomb” or “Quiescent/Low” chromatin states, but this is still no more than expected by chance. All  $P$ -values are Bonferroni corrected. States with enrichment  $P < 0.05$  are in bold. The corresponding fold enrichments are in Supplemental Figure 24.



**Figure 4.** Fold enrichments of mQTLs overlapping other QTLs and GWAS hits. mQTLs have greater enrichment for QTLs physically associated with DNA than they do for QTLs related to downstream traits only indirectly linked to DNA.

Across the different types of QTLs/GWAS that we compared to our mQTLs, we noticed an interesting trend: QTLs associated with the earliest steps of transcription (dsQTLs associated with chromatin accessibility) had the strongest overlap with mQTLs, while moving progressively towards phenotype, the enrichment decreased at each stage (Fig. 4). This suggests a buffering or dilution of mQTL effects at transcriptional and post-transcriptional levels, consistent with recent models for the attenuated effects of eQTLs on protein levels (Battle et al. 2014). As such, this could allow for plasticity in DNA methylation, with a small subset of methylation changes overcoming the many layers of gene expression regulation to ultimately affect organismal phenotypes.

## Discussion

The pooling approach introduced here represents a major advance in our ability to map *cis*-acting mQTLs genome wide. Our simulations demonstrated the substantial increase in power that our approach provides as a result of accounting for ASM. In an empirical test of our approach, we identified mQTLs in many functionally important sites that are not covered by microarrays, including TF binding sites, dsQTLs, CTCF-binding-QTLs, eQTLs, and GWAS hits. Thus, our approach overcomes two of the greatest weaknesses of microarrays—inability to detect ASM and limited coverage of CpGs—enabling us to detect thousands of novel mQTLs throughout the genome.

An additional advantage of our approach over microarrays and other individual-level assays is the minimization of between-sample experimental variability and batch effects, which are known to be important sources of error (Mostafavi et al. 2013). By pooling samples prior to bisulfite treatment, most potential sources of variability are eliminated. However, because our simulations did not account for these factors, we may have underestimated our increase in power over the traditional non-ASM method.

Several important advantages of the traditional approach of genotyping and using microarrays for DNA methylation measurement were also not captured by our simulations. For example, the

traditional approach can detect mQTLs acting at any distance away from their CpG targets, including *trans*-acting mQTLs. Because pooling requires the variant and CpG to be on the same sequenced DNA fragment, the distance is limited by the size of these fragments. (This includes the insert size in paired-end reads, so the distance can be many kilobases if large fragments are selected for sequencing, although a large variance in insert size can lead to fewer reads covering a given variant/CpG pair.) A previous study reported that the majority of “likely causal” mQTL variants are within 100 bp of the corresponding CpG (Banovich et al. 2014); therefore, our approach should be able to detect most of these causal variants, although it will have less power to detect additional mQTL variants that are in LD with the causal variants, since these can be much further away. As read lengths increase with new sequencing technologies, this limitation may be lessened.

Another limitation of our approach is that it excludes C/T and G/A SNPs from consideration because their genotypes cannot be disentangled from the effects of bisulfite conversion. However, because most SNPs (including C/T and G/A SNPs) are in strong linkage disequilibrium with other variants, their mQTL associations may still be measured via other “tag SNPs” (that are on the same read as the corresponding CpG), as is done in any QTL or GWAS study. Moreover, this limitation is shared by other methods for inferring ASM from individual-level bisulfite sequencing data.

Our approach also cannot account for population structure or any known covariates (e.g., gender or environmental differences) since the individual from which each read originated is unknown. In our study, this is probably not an issue because our individuals lack significant population structure, and the cell lines were grown in controlled laboratory conditions. Future applications of this approach would ideally use cohorts in which these issues are minimized, such as unstructured natural populations or controlled F2 crosses (as is ideal for any QTL study or GWAS) (Bush and Moore 2012).

Finally, our approach relies on detecting alleles directly from reads as well as detecting the positions of variants in the genome directly from reads if the data are from populations for which variant positions are not known. Detecting alleles from reads is not perfect due to sequencing errors and imperfect mapping. However, this is unlikely to have a substantial effect on our results because only errors at the variant positions would lead to an incorrect allelic assignment. Detecting variants is a more challenging problem (Nielsen et al. 2011; Li 2014), but it can be minimized by limiting the genotype calls to known variant positions, by sequencing deeply, or by excluding rare variants.

Although pooling individuals enables us to overcome limitations of previous approaches for mQTL mapping, we are not the first to use pooling for QTL mapping. Previous studies have used pooling for QTL or association mapping of individual traits. These approaches include bulk segregant analysis (BSA) (Michelmore and Paran 1991) and X-QTL mapping (Ehrenreich et al. 2010). They involve phenotyping many individuals for a specific trait followed by genotyping (via microarrays or sequencing) pools of individuals with either high or low trait values. QTLs are then identified as genetic variants with different allele frequencies between the two pools. These approaches have been gaining popularity because they, like ours, do not require individual-level genotyping, which is often the most laborious and expensive component of QTL mapping. In comparison to our method, BSA and X-QTL mapping can be applied to a much wider range of traits, including organismal-level traits. Two key advantages of our mQTL approach are that it can be applied to millions of

molecular-level traits in one pool (as opposed to a single trait for every two pools) and that it does not require any individual-level phenotyping.

We are also not the first to leverage allele-specific information for QTL mapping. Previous work has also used allele-specific information in ChIP-seq data to map QTLs associated with histone modification levels (McVicker et al. 2013) and to identify allele-specific histone modifications (Mikkelsen et al. 2007; Kasowski et al. 2013; Kilpinen et al. 2013). Unlike ChIP-seq, bisulfite sequencing results in approximately even sampling of all DNA in the sample, not just the DNA that is bound by a protein of interest. Therefore, our goal is not to identify the difference in read depth between alleles but rather the difference in the fraction of CpG methylation between alleles. Because these problems require distinct approaches, we proposed a novel approach for use in bisulfite sequencing studies.

Although our approach has many advantages over measuring genotypes and DNA methylation separately for each individual, there are ways to combine the approaches that would leverage ASM without sacrificing the ability to identify *trans*-acting and distal *cis*-acting mQTLs. For example, one could alter the step at which pooling takes place by creating a uniquely barcoded library for each sample prior to pooling and bisulfite treatment, thereby allowing each read to be assigned to its sample of origin. However, this would come at the cost of the additional experimental variability, effort, and expense associated with creating a separate library for each sample. Alternatively, even with no pooling, the extra power gained from allele-specific information in heterozygotes can be achieved by inferring both alleles and DNA methylation statuses directly from reads.

Large mQTL mapping studies have been limited by the effort and expense involved in data generation; our approach does not require individual-level data for either genotypes or DNA methylation, thereby significantly decreasing the barrier to mQTL mapping in any tissue or species. Although the number of mQTLs we identified was modest, this is primarily a limitation of our sequencing depth and not the method itself. As illustrated by our simulations, with 40× coverage, our power to detect even strong mQTLs is modest. As sequencing becomes less expensive, this pooling approach may help us achieve a comprehensive understanding of the relationship between genetic variation and DNA methylation, which will provide insight into traits such as evolutionary adaptations in many species and human diseases in many tissues. Moreover, in some modifications, the pooling framework we have introduced could also be applied to mapping QTLs for other molecular-level traits, such as TF binding and histone modifications. We anticipate that pooling will enable us to leverage sequencing technology to study the relationship between genetic and epigenetic variation in a much wider range of cell types and species than has previously been possible.

## Methods

### Simulations comparing the pooled and traditional association methods

In order to compare the power of identifying mQTLs using the pooled ASM method—sequencing on pooled samples and accounting for ASM—versus the traditional non-ASM method of associating genotypes with average methylation statuses, we performed simulations of a single theoretical variant-cytosine pair, varying several parameters of interest. We varied the effect

size (correlation between the variant's allele and the cytosine's methylation status), read depth (number of reads across all individuals), MAF and minor methylation status frequency (the minor methylation status is the cytosine methylation status with fewer reads), and number of individuals. We tested effect sizes ranging from 0.5 to 1.0; 20, 40, 80, 160, 320, 640, and 1280 reads across individuals; minor allele and methylation status frequencies 0.1, 0.2, 0.3, 0.4, and 0.5; and 25, 50, 100, 200, and 400 individuals. When assigning genotypes to individuals, we always assumed that our population is in Hardy-Weinberg equilibrium. For each parameter combination, we ran 10,000 simulations for the variant-cytosine pair.

We ran each simulation as if we were randomly selecting DNA fragments from a pool of cells. For the pooled data simulations, we sampled allele-methylation status pairs with replacement for each read. We used Fisher's exact test, implemented in MATLAB using Michael Boedigheimer's `fexact` function (<http://www.mathworks.com/matlabcentral/fileexchange/22550-fisher-s-exact-test>), to quantify the association. For the traditional non-ASM method simulations, we randomly assigned each read to an individual so that each individual had approximately (but not necessarily exactly) the same number of reads. For each individual, we used the genotype that we had initially assigned to it and computed the average methylation status, rounded to the nearest integer. We used a  $2 \times 3$  Fisher's exact test to compute the associations between genotype and average methylation status, which was implemented in MATLAB using Giuseppe Cardillo's `myfisher23` function (<http://www.mathworks.com/matlabcentral/fileexchange/15399-myfisher23>). For comparison, we also computed *P*-values using the *P*-value of the *F*-statistic for the regression that predicts methylation status as a function of allele/genotype and using the asymptotic *P*-value of the Pearson correlation between allele and methylation status; when we ran these simulations for the "traditional non-ASM method," we did not round the average methylation status for each individual. For each method, we estimated power as the fraction of simulations in which the variant-cytosine pair reached  $P < 0.001$ , the same *P*-value threshold that we used for identifying mQTLs in our real data. We then ran additional simulations as described in the Supplemental Methods.

### Whole-genome bisulfite sequencing library preparation

We pooled genomic DNA derived from 60 Yoruban LCLs (parental samples from 30 HapMap trios, purchased from Coriell). Forty-eight micrograms of this DNA was spiked with 240 ng unmethylated *cl857 Sam7 Lambda DNA* (Promega) to yield 0.5% W/W lambda DNA. The DNA was fragmented with a Covaris instrument (Covaris) in 50- $\mu$ L volumes (duty factor 10%; peak incident power 175,200 cycles per burst; 40-sec duration; 5.5°C–6.0°C), followed by end repair, adenylation, and adapter ligation using the TruSeq DNA LT Sample Prep Kit (Illumina) according to the manufacturer's instructions. Purification steps were performed using Agencourt Ampure beads (Beckman Coulter). All 24 indexed methylated adapters from TruSeq DNA LT Sample Prep Kit Sets A and B (Illumina) were used to construct the libraries in order to increase base complexity.

Adapter-ligated DNA of 400–500 bp was isolated by 2% agarose gel electrophoresis using low range ultra agarose (Bio-Rad) with SYBR Gold Nucleic acid gel stain (Invitrogen), and fractions were purified using MinElute Gel Extraction Kit (Qiagen). Two of the 24 libraries were reserved and not bisulfite converted for control purposes. Sodium bisulfite conversion was carried out on

each of the 22 remaining libraries using the EpiTect Bisulfite Kit (Qiagen #59104) according to the manufacturer's 2006 instructions, except the reaction mix incubation cycle from the Whole-Genome Bisulfite Sequencing for Methylation Analysis was used (Illumina, Part # 15021861 Rev. A) and consisted of 5 min at 95°C, 25 min at 60°C, 5 min at 95°C, 85 min at 60°C, 5 min at 95°C, 175 min at 60°C, three cycles consisting of 5 min at 95°C, and 180 min at 60°C, ending with a 20°C hold. Bisulfite-converted products were purified using the MinElute PCR Purification Kit (Qiagen).

Adapter-ligated, bisulfite-converted libraries were enriched using KAPA HiFi HotStart Uracil+ ReadyMix uracil-insensitive polymerase (D-mark Biosciences #KK2801). Thermocycler parameters for bisulfite-converted libraries consisted of 45 sec at 98°C; four cycles of 15 sec at 98°C, 30 sec at 65°C, 30 sec at 72°C, and ending with a 4°C hold. Adapter-ligated, non-bisulfite-converted control libraries were enriched using the reagents and protocol from the TruSeq DNA LT Sample Prep Kit (Illumina). Thermocycler parameters consisted of 30 sec at 98°C; four cycles of 10 sec at 98°C, 30 sec at 60°C, 30 sec at 72°C, and ending with a 4°C hold. PCR reaction products were purified using Agencourt Ampure beads (Beckman Coulter). Library validation was performed using the KAPA SYBR FAST Universal qPCR Library Quantification Kit (D-mark Biosciences, #KK4824) to measure the concentration of viable sequencing template molecules as well as the Agilent Bioanalyzer High Sensitivity DNA Assay (Agilent) to determine the size and distribution of the template molecules. Libraries were further concentrated using the MinElute PCR Purification Kit (Qiagen). Each noncontrol library was divided across eight sequencing lanes in three flowcells, and each control library was divided across four of these sequencing lanes, which were in two of the flowcells.

### Data processing

To obtain high-quality allele and methylation information from our bisulfite sequencing data, we trimmed reads, aligned reads to the genome, identified methylation statuses for cytosines and alleles for SNPs, and filtered our data according to various metrics (Towns et al. 2014). Details are described in the Supplemental Methods.

### Identifying methylation quantitative trait loci (mQTLs)

To find cytosines that are strongly associated with genetic variation, we combined variants in perfect linkage disequilibrium (LD), tested whether each association was significant, and used permutation tests to compute a false discovery rate (FDR). We associated every variant (including SNPs, insertions, and deletions) with every cytosine on the same read, creating a list of variant-cytosine pairs; thus, CpGs that never occurred on a read with a variant were not analyzed. When associating variants with CpGs, we combined cytosines on different strands of each CpG. For each cytosine, we combined associated variants whose genotypes in the 1000 Genomes AFR (The 1000 Genomes Project Consortium 2010) are perfectly correlated by replacing the allele of the variant that is later on the chromosome with the associated allele for the variant earlier on the chromosome (Supplemental Fig. 28). After combining SNPs, we computed the significance of the associations between variants' alleles and cytosines' methylation statuses for every variant-cytosine pair using Fisher's exact test (Fisher 1922), which we implemented using SciPy's Fisher exact test (Oliphant 2007). We removed all variant-cytosine pairs for which the num-

ber of reads with the minor allele or minor methylation status was so low that, given the total number of reads, the lowest  $P$ -value that could be achieved using Fisher's exact test was  $\geq 0.001$ . For each cytosine, we defined methylation quantitative trait loci (mQTLs) to be variants for which the association between the allele and the cytosine's methylation status had a Fisher's exact test  $P < 0.001$ ; this gave us an FDR of 0.0601. We computed the FDR by permuting the methylation statuses for each variant-cytosine pair 100 times, recomputing the  $P$ -value for each pair, reidentifying mQTLs, and averaging the number of mQTLs across permutations. The FDR is the ratio of the average number of mQTLs across permutations to the number of real mQTLs. To check for possible position-specific biases in our mQTL mapping, we intersected the variant-cytosine pairs for our real mQTL list with the variant-cytosine pairs for each permuted mQTL list; no variant-cytosine pair for a real mQTL occurred in more than two permuted lists, as expected.

### Computing mQTL enrichments

In order to test whether mQTLs associate with specific chromatin states or transcription factor (TF)-binding sites, we intersected our CpGs with mQTLs with chromatin states from GM12878 (Ernst and Kellis 2012; Kasowski et al. 2013) and our mQTLs with reproducible (IDR 2%) TF binding sites and histone modification regions from GM12878 (The ENCODE Project Consortium 2011, 2012; Li et al. 2011; Gerstein et al. 2012; Landt et al. 2012). If multiple data sets were available for a TF or histone modification, we used the data set with the largest number of peaks. We used the hypergeometric test, with the background defined as all variants in variant-cytosine pairs that were tested for being mQTLs. We corrected  $P$ -values using the Bonferroni correction (Shaffer 1995). For chromatin-state enrichments, we multiplied  $P$ -values by the number of possible chromatin states. For the TF binding site and histone modification enrichments in GM12878, we multiplied  $P$ -values by the number of TFs and histone modifications tested. In addition to computing  $P$ -values, we computed fold enrichments for chromatin states, TF binding sites, and histone modification regions.

To test whether mQTLs associate with open chromatin in various cell types, we intersected the variants we tested for mQTLs with Joseph Pickrell's list of SNPs in open chromatin regions in various cell types, which were downloaded from <https://github.com/joepickrell/1000-genomes> (Bernstein et al. 2010; The ENCODE Project Consortium 2012; Thurman et al. 2012). We used a hypergeometric test to evaluate the enrichment for mQTLs in variants overlapping open chromatin regions in each cell type. We then used a one-sided Wilcoxon rank-sum test to compare the mQTL enrichments for LCLs versus other cell types, in which our null hypothesis was that the median hypergeometric  $P$ -value for LCLs is not less than the median hypergeometric  $P$ -value for other cell types.

### Computing overlap between mQTLs, eQTLs, dsQTLs, CTCF-binding-QTLs, and GWAS hits

In order to evaluate whether mQTLs may be associated with gene expression, open chromatin, CTCF-binding, and organismal-level traits, we computed overlap between mQTLs and variants from expression QTLs (eQTLs), DNase hypersensitivity QTLs (dsQTLs), CTCF-binding-QTLs, and genome-wide association study (GWAS) data sets. To compute the enrichment of mQTLs in another molecular QTL study, we used a hypergeometric test, with the

background defined as all variants tested in both our study and the other study. In addition to computing  $P$ -values, we computed fold enrichments. The  $P$ -values and fold enrichments reported in the Results section are those for molecular QTLs and variants in perfect LD with molecular QTLs.

Because we do not know all of the SNPs in our study that have been tested in a GWAS, we applied a method similar to the one used by Lappalainen et al. (2013) for finding the enrichment of mQTLs in GWAS hits. We first created 10,000 null sets of variants that were tested for being mQTLs in our study, in which each variant  $i$  in a null set had an MAF within 0.001 of variant  $i$  in the set of mQTLs found in our study. Then we found the number of overlaps between the null set and GWAS hits. The  $P$ -value was defined as the fraction of null sets with at least as many overlaps with GWAS hits as mQTLs have with GWAS hits. We computed the fold enrichment for GWAS in the same way as the fold enrichment for molecular QTLs.

### Validation of mQTLs through bisulfite pyrosequencing

Bisulfite PCR-pyrosequencing assays were designed with PyroMark Assay Design 2.0 (Qiagen). The regions of interest were amplified by PCR using the HotstarTaq DNA polymerase kit (Qiagen) as follows: 15 min at 95°C (to activate the Taq polymerase), 45 cycles of 30 sec at 95°C, 30 sec at 58°C, and 30 sec at 72°C, and a 72°C 5-min extension step; primer sequences are listed in Supplemental Table 5. For pyrosequencing, a single-stranded DNA was prepared from the PCR product with the Pyromark Vacuum Prep Workstation (Qiagen), and the sequencing was performed using sequencing primers on a Pyromark Q96 MD pyrosequencer (Qiagen). The quantitative levels of methylation for each CpG dinucleotide were calculated with Pyro Q-CpG software (Qiagen).  $P$ -values for associations were the asymptotic  $P$ -values of the correlations between genotype and average methylation from the pyrosequencing assay. We performed pyrosequencing on the 60 individuals in our pool as well as on 30 additional individuals who are the offspring of those 60.

### Data access

Whole-genome bisulfite sequencing data generated in this study have been submitted to the NCBI Sequence Read Archive (SRA; <http://www.ncbi.nlm.nih.gov/sra>) under accession number SRP045408. Bisulfite pyrosequencing data are available in the Supplemental Material.

### Acknowledgments

We would like to thank M. Hirst, M. Moksa, R. White III, F. Krueger, A. Kundaje, N. Banovich, and X. Zhang for assistance. We would also like to thank J. Pritchard, C.S. Foo, M. Jones, and other members of the Fraser, Kobor, Koller, and Kundaje laboratories for useful discussions and feedback. I.M.K. is funded by the Department of Energy Computational Science Graduate Fellowship Program of the Office of Science and National Nuclear Security Administration in the Department of Energy under contract DE-FG02-97ER25308, the National Science Foundation Graduate Research Fellowship Program under grant number DGE-114747, and the Michael J. Flynn Stanford Graduate Fellowship. H.B.F. is an Alfred P. Sloan Fellow and a Pew Scholar in the Biomedical Sciences. M.S.K. is the Canada Research Chair in Social Epigenetics and a Senior Fellow of the Canadian Institute for Advanced Research. This work was supported by National Institutes of Health

grant number 1R21HG005240-01A1. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), through the San Diego Supercomputer Center Gordon Compute Cluster, which is supported by National Science Foundation grant number TG-MCB140156.

### References

- The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.
- Ambrosone CB, Young AC, Sucheston LE, Wang D, Yan L, Liu S, Tang L, Hu Q, Freudenheim JL, Shields PG, et al. 2014. Genome-wide methylation patterns provide insight into differences in breast tumor biology between American women of African and European ancestry. *Oncotarget* **5**: 237–248.
- Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y. 2014. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet* **10**: e1004663.
- Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. 2014. Impact of regulatory variation from RNA to protein. *Science* **347**: 664–667.
- Baylin SB, Herman JG. 2000. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet* **16**: 168–174.
- Baylin SB, Herman JG, Graff JR, Vertino PM, Issa JP. 1998. Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res* **72**: 141–196.
- Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* **12**: R10.
- Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A, et al. 2012. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet* **8**: e1002629.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**: 1045–1048.
- Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, et al. 2011. High density DNA methylation array with single CpG site resolution. *Genomics* **98**: 288–295.
- Bush WS, Moore JH. 2012. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**: e1002822.
- The Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* **490**: 61–70.
- Dayeh TA, Olsson AH, Volkov P, Almgren P, Rönn T, Ling C. 2013. Identification of CpG-SNPs associated with type 2 diabetes and differential DNA methylation in human pancreatic islets. *Diabetologia* **56**: 1036–1046.
- Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**: 390–394.
- De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, Eaton ML, Keenan BT, Ernst J, McCabe C, et al. 2014. Alzheimer's disease: early alterations in brain DNA methylation at *ANK1*, *BIN1*, *RHBDF2* and other loci. *Nat Neurosci* **17**: 1156–1163.
- Ding Z, Ni Y, Timmer SW, Lee BK, Battenhouse A, Louzada S, Yang F, Dunham I, Crawford GE, Lieb JD, et al. 2014. Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet* **10**: e1004798.
- Drong AW, Nicholson G, Hedman ÅK, Meduri E, Grundberg E, Small KS, Shin SY, Bell JT, Karpe F, Soranzo N, et al. 2013. The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. *PLoS One* **8**: e55923.
- Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, Gresham D, Caudy AA, Kruglyak L. 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature* **464**: 1039–1042.
- The ENCODE Project Consortium. 2011. A user's guide to the Encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**: e1001046.
- The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.
- Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.
- Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schübeler D. 2013. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet* **9**: e1003994.

- Fisher RA. 1922. On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P. *J R Stat Soc* **85**: 87.
- Fraser HB, Lam LL, Neumann SM, Kober MS. 2012. Population-specificity of human DNA methylation. *Genome Biol* **13**: R8.
- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.
- Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci* **89**: 1827–1831.
- Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, Cox NJ, Gershon ES, Kelsoe JR, Greenwood TA, Nievergelt CM, et al. 2012. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder susceptibility variants. *Mol Psychiatry* **3**: 340–346.
- Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**: 91–100.
- Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, et al. 2010. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* **6**: e1000952.
- Grundberg E, Meduri E, Sandling JK, Hedman ÅK, Keildson S, Buil A, Busche S, Yuan W, Nisbet J, Sekowska M, et al. 2013. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet* **93**: 876–890.
- Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, et al. 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**: e00523.
- Heyn H. 2014. A symbiotic liaison between the genetic and epigenetic code. *Front Genet* **5**: 113.
- Heyn H, Moran S, Hernando-Herrera I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L, et al. 2013. DNA methylation contributes to natural human variation. *Genome Res* **23**: 1363–1372.
- Irier HA, Jin P. 2012. Dynamics of DNA methylation in aging and Alzheimer's disease. *DNA Cell Biol* **31**: S42–S48.
- Irizarry RA, Wu H, Feinberg AP. 2009. A species-generalized probabilistic model-based definition of CpG islands. *Mamm Genome* **20**: 674–680.
- Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**: 118–127.
- Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**: 484–492.
- Jones MJ, Fejes AP, Kober MS. 2013. DNA methylation, genotype and gene expression: who is driving and who is along for the ride? *Genome Biol* **14**: 5–7.
- Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV, et al. 2013. Extensive variation in chromatin states across humans. *Science* **342**: 750–752.
- Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, Li K, Murty VV, Schupf N, Vilain E, et al. 2008. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat Genet* **40**: 904–908.
- Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, et al. 2013. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**: 744–747.
- Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, Kober MS. 2012. Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci* **109**: 17253–17260.
- Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**: 1813–1831.
- Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, González-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–511.
- Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**: 733–739.
- Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**: 2843–2851.
- Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**: 1752–1779.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**: 315–322.
- Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, et al. 2013. Global epigenomic reconfiguration during mammalian brain development. *Science* **341**: 1237905.
- Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, Howard TD, Hawkins GA, Cui W, Morris J, et al. 2013. Methyloomics of gene expression in human monocytes. *Hum Mol Genet* **22**: S065–S074.
- Lunnon K, Smith R, Hannon E, De Jager PL, Srivastava G, Volta M, Troakes C, Al-Sarraj S, Burrage J, Macdonald R, et al. 2014. Methyloomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nat Neurosci* **17**: 1164–1170.
- McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, Lewellen N, Myrthil M, Gilad Y, Pritchard JK. 2013. Identification of genetic variants that affect histone modifications in human cells. *Science* **342**: 747–749.
- Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**: S868–S877.
- Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**: 766–770.
- Michelmore R, Paran I. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci* **88**: 9828–9832.
- Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**: 553–560.
- Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, Myers J, Godley LA, Dolan ME, Zhang W. 2013. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics* **194**: 987–996.
- Mokry LE, Ahmad O, Forgetta V, Thanassoulis G, Richards JB. 2014. Mendelian randomisation applied to drug development in cardiovascular disease: a review. *J Med Genet* **52**: 71–79.
- Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, Koller D. 2013. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One* **8**: e68141.
- Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**: 443–451.
- Oliphant TE. 2007. Python for scientific computing. *Comput Sci Eng* **9**: 10–20.
- Pease M, Ling C, Mack WJ, Wang K, Zada G. 2013. The role of epigenetic modification in tumorigenesis and progression of pituitary adenomas: a systematic review of the literature. *PLoS One* **8**: e82619.
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**: 768–772.
- Schmitz RJ, He Y, Valdés-López O, Khan SM, Joshi T, Urich MA, Nery JR, Diers B, Xu D, Stacey G, et al. 2013a. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res* **23**: 1663–1674.
- Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, et al. 2013b. Patterns of population epigenomic diversity. *Nature* **495**: 193–198.
- Shaffer JP. 1995. Multiple hypothesis testing. *Annu Rev Psychol* **46**: 561–584.
- Shi J, Marconett CN, Duan J, Hyland PL, Li P, Wang Z, Wheeler W, Zhou B, Campan M, Lee DS, et al. 2014. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat Commun* **5**: 3365.
- Shoemaker R, Deng J, Wang W, Zhang K. 2010. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res* **20**: 883–889.
- Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**: 204–220.
- Smith AK, Kilaru V, Kocak M, Almlil LM, Mercer KB, Ressler KJ, Tylavsky Fa, Conneely KN. 2014. Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* **15**: 145.
- Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**: 500–507.
- Sun Z, Chai H, Wu Y, White WM, Donkena KV, Klein CJ, Garovic VD, Therneau TM, Kocher JP. 2011. Batch effect correction for genome-

- wide methylation data with Illumina Infinium platform. *BMC Med Genomics* **4**: 84.
- Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr AR, Deaton A, Andrews R, James KD, et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**: 1082–1086.
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.
- Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD, et al. 2014. XSEDE: accelerating scientific discovery. *Comput Sci Eng* **16**: 62–74.
- Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. 2014. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol* **15**: R37.
- Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**: 1095–1106.
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorf L, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**: D1001–D1006.
- Wiench M, John S, Baek S, Johnson TA, Sung MH, Escobar T, Simmons CA, Pearce KH, Biddie SC, Sabo PJ, et al. 2011. DNA methylation status predicts cell type-specific enhancer activity. *EMBO J* **30**: 3028–3039.
- Wolf SF, Jolly DJ, Lunnen KD, Friedmann T, Migeon BR. 1984. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proc Natl Acad Sci* **81**: 2806–2810.
- Wrzodek C, Büchel F, Hinselmann G, Eichner J, Mittag F, Zell A. 2012. Linking the epigenome to the genome: correlation of different features to DNA methylation of CpG islands. *PLoS One* **7**: e35327.
- Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP. 2010. Redefining CpG islands using hidden Markov models. *Biostatistics* **11**: 499–514.
- Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**: 816–831.
- You JS, Kelly TK, De Carvalho DD, Taberlay PC, Liang G, Jones PA. 2011. OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. *Proc Natl Acad Sci* **108**: 14497–14502.
- Yousefi P, Huen K, Schall RA, Decker A, Elboudwarej E, Quach H, Barcellos L, Holland N. 2013. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics* **8**: 1141–1152.
- Zhang X, Moen EL, Liu C, Mu W, Gamazon ER, Delaney SM, Wing C, Godley LA, Dolan ME, Zhang W. 2014. Linking the genetic architecture of cytosine modifications with human complex traits. *Hum Mol Genet* **22**: 5893–5905.
- Zhi D, Aslibekyan S, Irvin MR, Claas SA, Borecki IB, Ordovas JM, Absher DM, Arnett DK. 2013. SNPs located at CpG sites modulate genome-epigenome interaction. *Epigenetics* **8**: 802–806.
- Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, et al. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**: 477–481.

Received August 29, 2014; accepted in revised form April 17, 2015.