Genome Biology

RESEARCH Open Access

Population-specificity of human DNA methylation

Hunter B Fraser[1*], Lucia L Lam[2,3], Sarah M Neumann[2,3] and Michael S Kobor[2,3*]

Abstract

Background: Ethnic differences in human DNA methylation have been shown for a number of CpG sites, but the genome-wide patterns and extent of these differences are largely unknown. In addition, whether the genetic control of polymorphic DNA methylation is population-specific has not been investigated.

Results: Here we measure DNA methylation near the transcription start sites of over 14, 000 genes in 180 cell lines derived from one African and one European population. We find population-specific patterns of DNA methylation at over a third of all genes. Furthermore, although the methylation at over a thousand CpG sites is heritable, these heritabilities also differ between populations, suggesting extensive divergence in the genetic control of DNA methylation. In support of this, genetic mapping of DNA methylation reveals that most of the population specificity can be explained by divergence in allele frequencies between populations, and that there is little overlap in genetic associations between populations. These population-specific genetic associations are supported by the patterns of DNA methylation in several hundred brain samples, suggesting that they hold *in vivo* and across tissues.

Conclusions: These results suggest that DNA methylation is highly divergent between populations, and that this divergence may be due in large part to a combination of differences in allele frequencies and complex epistasis or gene × environment interactions.

Background

In multicellular organisms, the great diversity of cell types is maintained by mitotically heritable differences in gene expression, which are in part regulated by epigenetic mechanisms [1]. These include histone modifications, histone variants, RNA-based mechanisms, and DNA methylation [2]. The latter is perhaps the best understood component of the epigenetic machinery [3] and in somatic cells occurs almost exclusively on cytosine residues in the context of CpG dinucleotides [4]. While CpGs are underrepresented across the human genome, they are enriched at the majority of gene promoters, forming regions known as CpG islands that can regulate the expression of neighboring genes [4]. DNA methylation is not only closely linked to tissue-specific gene expression, but also to a number of intriguing biological phenomena such as X-chromosome inactivation in females, allele-specific expression of imprinted genes, aging, and cancer [5].

An emerging aspect of epigenetics is its role at the interface between the environment and the genome [6]. Although DNA methylation is a very stable epigenetic mark, numerous environmental influences have been associated with variation in DNA methylation as well as other epigenetic marks [2,6]. These include nutritional factors, exposure to environmental pollutants, and social environment. It is this plasticity that underlies much of the potential contribution of DNA methylation to multifactorial diseases and complex phenotypes [7]. However, the fundamental biology of the epigenome poses some challenges to testing this attractive concept. For example, most primary material available from human populations consists of mixtures of different cell types with distinct epigenomes, making it difficult to specifically assess the association of epigenetic changes with environmental exposure and phenotype. To address the role of epigenetics in common disease, it is important to understand the nature of epigenetic variation in the context of genetically well-characterized pure cell populations.

* Correspondence: hbfraser@stanford.edu; msk@cmmt.ubc.ca
[1]Department of Biology, Stanford University, Stanford, CA 94305, USA
[2]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, V6T 1Z3, Canada
Full list of author information is available at the end of the article

Recent advances in high-throughput technologies for measuring DNA methylation have allowed the patterns of methylation to be characterized throughout the human genome [8-15]. Comparing these results between twins has revealed that methylation at some CpG sites can be heritable [14,15], and combining them with genotype data has led to the discovery of hundreds of methylation-associated SNPs, or 'mSNPs', in brain tissue [11,12] as well as cell lines [13]. However, the question of whether the effects of mSNPs on DNA methylation levels and heritability differ between human populations has not been addressed. Quantifying such population specificity is important for our understanding of the genetic architecture of the epigenome, as well as its plasticity during human evolution.

Results

To compare DNA methylation between human populations, we utilized lymphoblastoid cell lines (LCLs) from the HapMap project [16], which have been extensively genotyped and previously employed to study the population specificity of gene expression levels [17-19]. Although LCLs can acquire changes in gene expression and DNA methylation during transformation and cell culture [20,21], it has been shown that the inter-individual variation - which is what is relevant for the current work - is nearly always conserved (at least for gene expression) [21]. Our initial study set consisted of 30 family trios (mother/father/offspring) of Northern European ancestry (abbreviated CEU), and 30 trios of Yoruban (West African) ancestry (abbreviated YRI). These 180 cell lines were grown in identical conditions and their genomic DNA was subjected to quantitative bead-array-based DNA methylation analysis at 27, 578 CpG sites near the transcription start sites of 14, 495 genes (Materials and methods). Although an average of approximately two CpG sites near each transcription start site does not directly measure most of the methylation in regulatory regions, the fact that sites separated by under approximately 1 kb show highly correlated methylation [9,10] suggests that our data may actually capture the majority of methylation information near transcription start sites - similar to the effect of linkage disequilibrium (LD) between genetic variants in genome-wide association studies (though there is no guarantee that the most relevant sites will be in 'methylation LD' with the CpG sites we measure). The 1, 092 sites on the × and Y chromosomes were excluded from all analyses to eliminate gender effects, leaving 26, 486 autosomal sites in 13, 890 genes (in which no significant sex specificity was observed; Figure S1 in Additional file 1).

The resulting data revealed a wide range of within-population variability in the methylation of individual CpG sites (Figure 1a), consistent with previous work [11-13]. Across all sites, the average correlation of methylation profiles between individuals (mean r^2 = 0.78 for CEU, 0.86 for YRI) was far lower than that of technical replicates (r^2 > 0.99 for all six replicate pairs), indicating that most of the variability was biological, and not technical. In addition, we replicated results for two variable sites in all 180 samples by pyrosequencing bisulfite-treated DNA. This showed excellent concordance with our array-based results (r^2 = 0.88 for *IGSF2* and 0.94 for *PLSCR2*; Figure 1b), suggesting that the array data provide accurate quantification of DNA methylation levels.

In addition to the variation within each population, we observed extensive differences in the DNA methylation patterns between populations (for example, *FLJ32569* in Figure 1a). To quantify this population specificity, we calculated the number of CpG sites with methylation differing between populations, using the nonparametric Wilcoxon test. We found a substantial fraction differing between the populations (Figure 1c): at nominal P < 0.01, 8, 475 sites differed between populations (32.0% of sites; false discovery rate (FDR) = 3.1%), and 5, 654 sites remained significant at P < 0.001 (21.4% of sites; FDR = 0.5%; Figure S2 in Additional file 1). Thus, the methylation of approximately 30% of the CpG sites we studied - representing over a third of the genes assayed - differed between populations (this degree of population specificity is similar to that of gene expression levels in the same cell lines; Figure S3 in Additional file 1). However, these population-level differences tended to be small in magnitude, with only 1, 033 sites (3.9%) differing by an average of over 10% methylation, and 3, 695 sites (14.0%) differing by over 5%. Perhaps because of their small magnitudes, differences in DNA methylation explained very little of the variation in gene expression levels between populations that has been previously reported [17-19] (Supplemental text and Figure S4 in Additional file 1), consistent with previous findings that inter-individual variation in DNA methylation explains almost none of the variation in gene expression [12,13].

These subtle but extensive epigenetic differences between populations could have genetic or environmental underpinnings - or a combination of both. To assess the role of both common and rare genetic variants in determining DNA methylation patterns, we estimated the contribution of additive genetic variation (known as narrow-sense heritability, or h^2) to the methylation of each CpG site in each population by measuring the correlation in methylation levels between parents and their offspring (Figure 2a; Materials and methods). We observed heritable methylation at approximately 762 CpG sites in CEU and 930 sites in YRI (Figure 2b), suggesting that genetic control of polymorphic methylation is fairly common - though slightly less heritable than

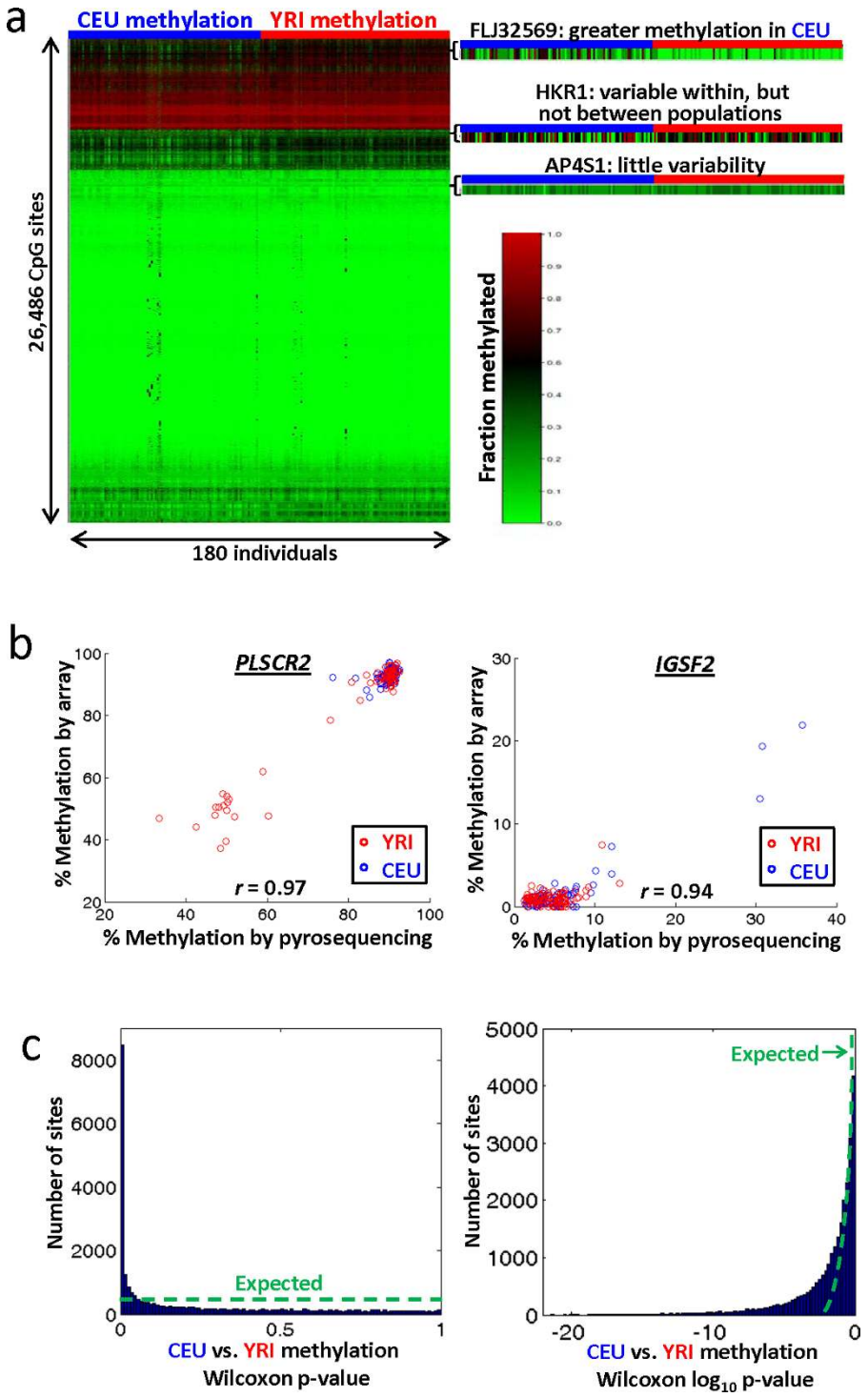

Figure 1 Population-specificity of DNA methylation. **(a)** Heatmap of the clustered methylation data set. Three representative cases are magnified: a site with a clear population difference; a site showing within- but not between-population variability; and a site with little variability within or between populations. **(b)** We performed pyrosequencing as an independent means to measure methylation of two CpG sites (*IGSF2*, chromosome 1, base 117345939; *PLSCR2*, chromosome 3, base 147696535) in our 180 samples. The agreement validates the accuracy of our microarray data. **(c)** The methylation of many sites differs between CEU and YRI. We performed the nonparametric Wilcoxon test to identify CpG sites differing in methylation between populations. The *P*-values are skewed towards small values, as shown by comparing to the expected uniform distribution on either a linear (left) or log (right) scale.

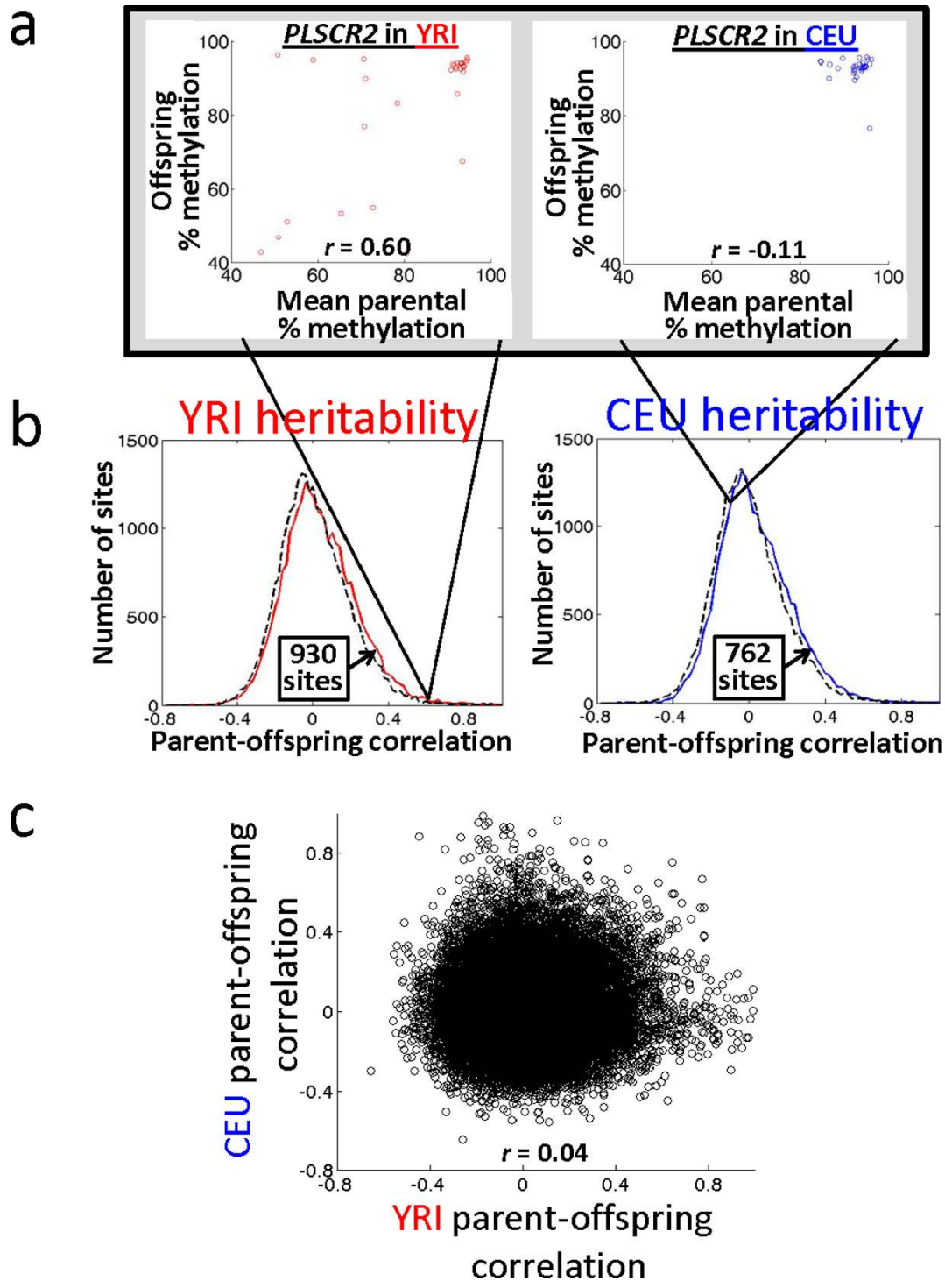

Figure 2 Population specificity of DNA methylation heritability. **(a)** An example of a CpG site (near *PLSCR2*: chromosome 3, base 147696535) whose methylation is heritable in YRI, but not CEU, as assessed by the similarity of average parental methylation to their offspring methylation (each point represents one family trio). **(b)** Histograms comparing the observed distribution of per-site heritabilities to a typical randomized distribution (numbers in the text are based on 1, 000 randomizations; Materials and methods). The greater number of sites at high heritabilities in the real data compared to random (arrows) is an estimate of the number of heritable sites we can detect in each population. **(c)** No similarity between heritabilities in each population (Pearson's r^2 = 0.002; each point is a CpG site).

gene expression levels in the same cell lines (Figure S5 in Additional file 1). Given our limited power to detect weakly heritable DNA methylation, these numbers are likely to be substantial underestimates of the true extent of heritability.

Considering the overall genetic similarity among human populations [16,22], we expected the patterns of heritability in CEU and YRI to be similar. Surprisingly, we found almost no correlation between them (r^2 = 0.002; Figure 2c). This is similar to agreement in h^2 for gene expression levels in the same cell lines (Figure S6 in Additional file 1). We did not find any evidence for complex inheritance patterns - such as dominance, maternal-biased, or paternal-biased inheritance of DNA methylation - that could affect heritability (Supplemental text in Additional file 1).

Differences in heritability between populations could have many causes. h^2 is defined as the ratio of a trait's additive genetic variance to its total variance in a population; factors that can affect this ratio include changes in the additive genetic variance (for example, differing allele frequencies), non-additive (gene × gene, or GxG) genetic variance, environmental variance, and gene × environment (GxE) interaction variance [23]. In addition, limited statistical power could restrict the accuracy of our heritability estimates (Supplemental text and Figure S7 in Additional file 1). Although we were not able to rule out any of these potential factors, the extensive DNA sequence data available for these samples do allow us to test the contributions of two types of divergence that may contribute to the population-specific DNA methylation levels, and their heritabilities.

One type of divergence that may affect DNA methylation levels and heritabilities is a difference in the CEU/YRI allele frequencies at genetic variants that influence methylation. In particular, lower minor allele frequency at such a variant reduces the population-level genetic variation affecting a site's methylation, thus reducing h^2. To test how much of our observed population specificity can be explained in this way, we first identified the 'local' SNP (within 100 kb of the CpG) most strongly associated with each CpG's methylation across all 180 samples from both populations (although genetic associations in ethnically heterogeneous cohorts such as this can reflect population stratification, it is appropriate for our current goal). We then included this single SNP genotype in a multiple regression analysis to assess whether genotype or population was a stronger predictor of methylation at each site. Among the 5, 654 CpG sites differing between populations at Wilcoxon $P <$ 0.001 (discussed above), we found that 3, 131 (55.4%) were more strongly associated with a local SNP genotype than with population, implying that common (and likely *cis*-acting) genetic variants can explain over half of the population specificity we observed. This result also indicates that most of the population specificity is unlikely to be due to any type of cell line artifacts, since these would not correlate with individual SNP genotypes.

The second type of divergence we tested concerned complex GxG or GxE interactions: if a genetic variant is present in two populations, but affects DNA methylation in only one, then that variant must genetically interact with other variants and/or the environment. Such interactions can decrease heritability by increasing the population-level variance in DNA methylation (the denominator of h^2) without affecting the additive genetic variance (the numerator). To perform this analysis, we needed to identify SNPs associated with the methylation of individual CpG sites separately in each population, and then compare the lists to one another.

Three previous studies of genome-wide DNA methylation have mapped SNPs whose genotype correlates with the methylation of a CpG site, termed 'mSNPs' [11-13]. Because mSNPs are highly enriched close to their target CpG sites [11-13], we performed a 'local' association analysis between methylation at each CpG site with all HapMap SNPs within 100 kb, separately for each population. These local mSNP associations can arise from either true (likely *cis*-acting) genetic associations, or genetic variants that disrupt hybridization of the bead-array probes in some individuals, leading to spurious associations (analogous to issues in eQTL mapping [24]). Using recent and essentially complete catalogs of common genetic variants in each [22], we identified all probes overlapping variants present in the 1000 Genomes samples (2, 734 probes in CEU, and 3, 923 probes in YRI; Table S1 in Additional file 1). We observed a 2.6-fold higher frequency of mSNPs for these probes compared to probes not disrupted by SNPs, implying a high rate of spurious associations (re-analysis of previously reported brain mSNPs [11,12] suggests a similarly high rate of spurious associations in those studies). Therefore, we removed these probes from our analysis (these sites did not have a higher level of heritability or population differentiation, so were not excluded from those analyses; Supplemental text and Figure S8 in Additional file 1).

After excluding the potentially problematic probes, we identified 49 mSNPs in CEU and 86 in YRI (genotype versus methylation level $r > 0.6$; FDR of 37% and 28%, respectively), each explaining 36 to 92% of the variance in DNA methylation at the associated site (Figure 3a). We note that these numbers are not directly comparable to previous studies [11,12] that included CpG probes that may contain SNPs, since including probes overlapping SNPs in our analysis increases the number of (apparent) mSNPs while decreasing the FDR. Restricting

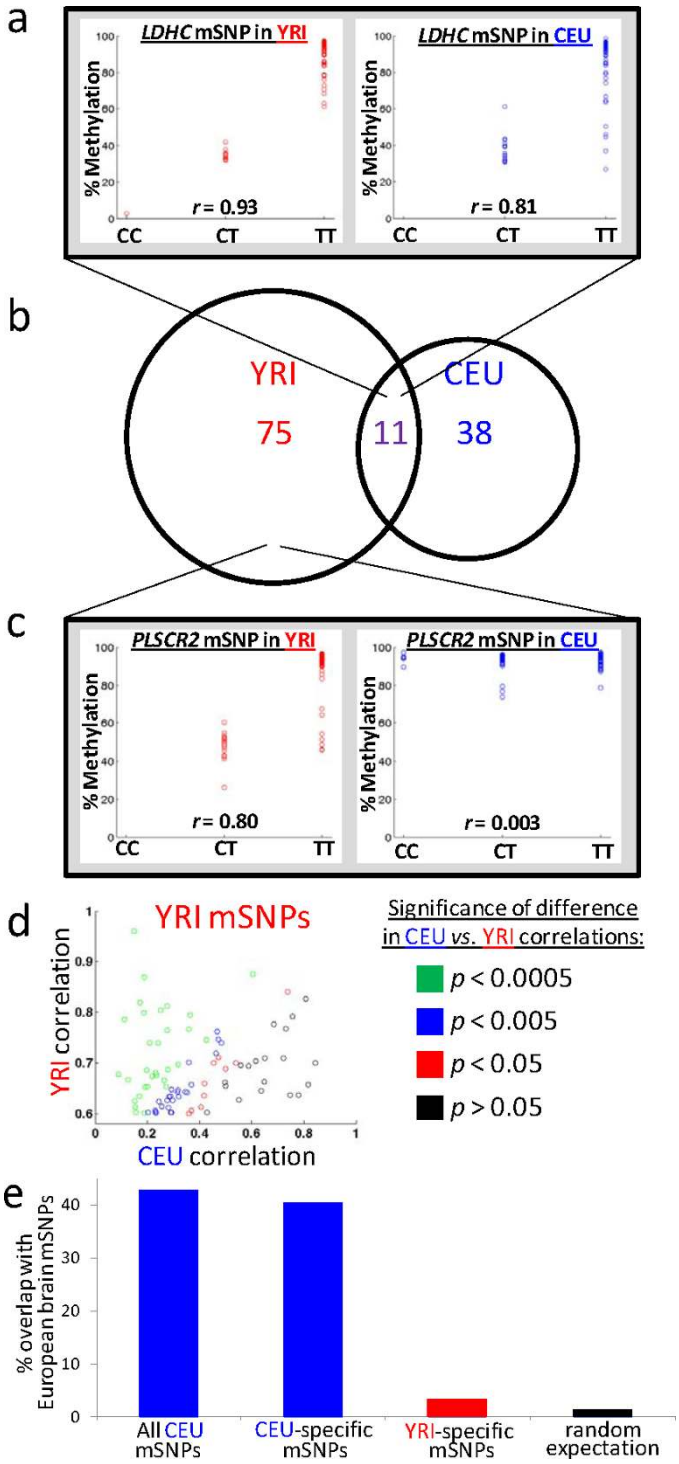

Figure 3 Population specificity of mSNPs. **(a)** An example of an mSNP (between a CpG site near *LDHC* (chromosome 11, base 18390591), and rs2643856) that is found in both YRI and CEU. In both cases the T allele is associated with higher methylation. **(b)** Venn diagram of the overlap among CpG sites associated with an mSNP in YRI and/or CEU. Five CEU sites and eight YRI sites were excluded from the overlap analysis because they overlapped a SNP in the other population. **(c)** Example of an mSNP (between a CpG site near *PLSCR2* (chromosome 3, base 147696535) and rs12489924) that is found in YRI but not CEU. No other SNPs in CEU within 100 kb of the CpG are associated with methylation at the site ($r < 0.25$ for all), indicating that the difference is unlikely to be due to differing LD between rs12489924 and the causal variant. **(d)** Scatter plot of all 86 YRI mSNPs, showing the strongest association found for that site in each population. Points are colored according to the significance of the difference in the associations within each population; most mSNP association strengths are significantly ($P < 0.005$) different between populations. The same plot for CEU mSNPs is shown in Figure S10 in Additional file 1. **(e)** Overlap of LCL mSNPs with brain mSNPs from two studies of European populations (similar to CEU). Both all CEU mSNPs and CEU-specific mSNPs show similar overlap of 40 to 42%, which is thus a minimum estimate for the extent of mSNPs shared between LCLs and brain. However, YRI-specific mSNPs show only 3.2% overlap, not significantly different from the 1.2% expected from any random set of CpG sites.

the CpG sites to only those with heritable methylation ($h^2 > 0.2$) decreased the FDR substantially (24 mSNPs at 8.6% FDR in CEU; 55 mSNPs at 4.7% FDR in YRI), providing a high-confidence list of mSNPs (Table 1; Table S2 in Additional file 1), as well as evidence supporting our heritability estimates in each population. Our high-confidence YRI mSNP list overlapped the mSNPs from a previous study of YRI LCL mSNPs [13] over 50-fold more than expected by chance (Supplemental text in Additional file 1). The vast majority of our mSNPs did not coincide with eSNPs (SNPs associated with gene expression levels; Supplemental text in Additional file 1), in agreement with previous work [13], suggesting that most do not impact gene expression levels in standard LCL culture conditions. None of these mSNPs affected methylation in known imprinted regions, and there was no enrichment for Gene Ontology categories or KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways among the genes associated with either population's mSNPs.

To test our mSNP mapping accuracy, we performed bisulfite Sanger sequencing at one mSNP locus (*RNF186*; Table 1) on 55 individual DNA molecules from six samples (three CEU and three YRI; Figure S9 in Additional file 1). Each individual's average methylation level at a particular CpG site (cg09195271) agreed with our array-based results ($r^2 = 0.74$), recapitulating the association between this site's methylation and the genotype of a nearby SNP (rs3806308): individuals with a CC genotype had the lowest average methylation (4/27 DNA molecules methylated = 14.8%), CT was intermediate (5/18 = 27.8% methylated), and TT had the highest (8/10 = 80% methylated). Interestingly, the methylation at six additional CpG sites in between rs3806308 and the target CpG did not correlate with the SNP genotype, indicating site-specific control of methylation, and not a more general regional effect.

Comparing our complete catalogs of mSNPs from each population, we found little overlap between them, or in the DNA methylation sites associated with mSNPs:

Table 1 High-confidence mSNPs in CEU

Gene	Chromosome	CpG position	mSNP	Perccentage CEU variance explained	Perccentage YRI variance explained	CEU h^2	YRI h^2	Brain mSNP?
TTC13	1	229182620	rs7545429	71.3	49.0	0.41	0.64	No
MGC3207	19	13736014	rs371671	68.8	27.2	0.60	0.35	Yes
PPP4R2	3	73128376	rs9816164	66.7	43.2	0.51	0.23	Yes
LDHC	11	18390591	rs11601413	65.4	86.5	0.55	0.68	Yes
RNF186	1	20015084	rs3806308	65.1	68.3	0.41	0.50	No
FLJ32569	1	204085874	rs823080	58.5	**4.5**	0.28	0.05	Yes
NDUFAF2	5	60275337	rs162244	57.4	62.6	0.26	0.49	No
PCGF3	4	689950	rs2242234	57.2	**19.9**	0.47	-0.10	No
LTA	6	31648435	rs2516390	55.9	40.5	0.48	0.24	No
IGSF2	1	117345939	rs12130298	52.6	**10.0**	0.96	-0.19	No
GSTM5	1	110056139	rs4970776	52.4	**12.1**	0.55	0.14	Yes
FLJ32569	1	204085802	rs823080	50.4	**3.7**	0.49	0.08	Yes
ASCIZ	16	79627243	rs16954698	47.8	**9.6**	0.24	-0.12	No
TACSTD2	1	58815787	rs1109896	42.2	50.4	0.29	0.49	No
HLA-C	6	31347299	rs6457375	42.1	44.0	0.24	0.61	Yes
HLA-DRB5	6	32606582	rs9271586	42.0	28.2	0.32	0.42	No
LYCAT	2	30523367	rs829650	40.8	52.4	0.75	0.64	Yes
PARK2	6	163069159	rs13218900	40.4	41.6	0.21	0.03	No
ITPR1	3	4510075	rs304075	39.4	**7.6**	0.21	-0.07	No
PSMD5	9	122644335	rs12343516	39.4	35.1	0.53	0.11	Yes
BTN3A2	6	26472772	rs2393667	38.1	**14.9**	0.22	0.31	Yes
RAPGEF3	12	46439111	rs3759407	37.2	**6.8**	0.71	-0.17	No
FAM83A	8	124264314	rs16898095	36.3	76.5	0.27	0.71	No
CRIP2	14	105011436	rs4983346	36.1	**3.6**	0.46	0.04	No

The 24 mSNP-CpG site pairs where > 36% of the variance in CEU methylation is explained by the mSNP genotype, and $h^2 > 0.2$. When more than one SNP was tied for the strongest association (due to perfect LD), one was chosen randomly. The YRI association strength is for the top local (within 100 kb) mSNP association for the same CpG site. In bold are YRI associations that explain < 20% of the variance in YRI methylation, indicating a high-confidence set of CEU-specific associations. For brain mSNPs, the intersection of *cis*-acting mSNP lists used by the authors of each original study [11,12] was used. YRI mSNPs are listed in Table S2 in Additional file 1.

only 11 CpG sites (8.9% of the mSNP-associated sites) were present in both of our medium-confidence lists (Figure 3a-c). This lack of overlap parallels the extensive population specificity of both methylation levels (Figure 1c) and their heritabilities (Figure 2c). Sites with population-specific mSNPs also tended to have population-specific heritabilities (Table 1, entries in bold; and see *PLSCR2* in Figure 2a and 3b), suggesting that the mSNPs we detect are a major source of the heritability of their target sites' methylation.

Three factors could contribute to a lack of overlap between mSNPs from each population: low power, differing LD/allele frequencies, and true population-specific effects of genetic variation on methylation. We found that neither low power nor differing LD/allele frequencies could account for most of the population specificity we observed (Supplemental text in Additional file 1), suggesting that many mSNPs exert population-specific effects on DNA methylation. Such population specificity can only be explained by interactions between the mSNPs and other genetic variants, and/or the environment (see Discussion).

Comparing our mSNP catalogs to previously reported mSNPs from brain allows us to test the generality of the observed population specificity in an independent cohort and tissue. Among our CEU mSNPs, 42% (10/24; Figure 3e) were previously observed in both of two brain mSNP catalogs that utilized cohorts of European ancestry [11,12] (Table 1), indicating that these associations are shared across tissues. A similar fraction (4/10, 40%; Figure 3e; Table 1, entries in bold) of the subset of high-confidence mSNPs observed only in CEU (not YRI) were also seen in brain. A key prediction of our results is that mSNPs found only in YRI should not be observed in the European brain samples if they are truly population specific. In support of this, only 1/32 (3.1%; Figure 3e; Table S1 in Additional file 1) of YRI-specific mSNPs were seen in European brain (not significantly different than the 1.2% expected by chance). This lack of overlap is unlikely to be due to potential artifacts of long-term cell culture, since the CEU cell lines are decades older than the YRI, which would tend to act against the trend we observed. Therefore, we conclude that the population specificity we discovered is recapitulated *in vivo*, as well as across tissues.

Discussion

Our results demonstrate extensive population specificity in DNA methylation profiles near transcription start sites. We observed these differences at three levels: the extent of DNA methylation, its heritability, and its association with specific genetic variants (mSNPs). We attribute most of these differences to two main factors: population-specific allele frequencies of genetic variants affecting DNA methylation, and complex GxG or GxE interactions.

Although *in vitro* artifacts are always a concern when using cell lines - and in particular LCLs, which have been shown to have some methylation differences compared to blood [20,21] - our results are unlikely to be driven by these effects, for three main reasons. First, unlike some previous studies of population-level differences in these cell lines [17,25], we processed samples in a randomized design, to eliminate the possibility of batch effects influencing our estimates of population specificity. Second, we found most of the population-specific DNA methylation to be explained by local genetic variants, ruling out any type of cell line artifact as an alternative explanation. Third, and most importantly, our population-specific mSNPs are supported by comparison to two studies of brain mSNPs in cohorts of European ancestry: 40% of our CEU-specific mSNPs overlap with both of these previous studies, whereas only 3.1% of YRI-specific mSNPs do, despite our expectation that the much older CEU LCLs would be more likely to have accumulated abnormalities in DNA methylation [20]. Together, these lines of evidence strongly suggest that our results apply *in vivo* and across tissues.

A variant that is present in two populations, but affects DNA methylation in only one, can only be explained by complex genetic interactions. These interactions could involve the environment (GxE), epistasis with other variants (GxG), or both. For example, some genetic variants have an observable effect on DNA methylation only in the presence of a sufficient quantity of methyl donors [26], which could differ between Yorubans and European-Americans as a result of diet or other factors (though methylation differences due to GxE interactions would have to be preserved during the creation and culturing of the LCLs). Even with such interactions causing differentiation between populations, genetic effects could be entirely additive within populations, consistent with our observation of heritable DNA methylation at many sites.

Divergence in the genetic underpinnings of DNA methylation (as evidenced by the population-specific mSNPs) would be expected to result in differing heritabilities and methylation levels, consistent with our results. Although we cannot provide an accurate estimate of exactly how much of the population-specific DNA methylation we observed is due to population-specific mSNPs, it is likely to be a substantial fraction once mSNPs of small effect (which could not be detected here due to our limited sample size) are accounted for.

Conclusions

As DNA methylation is an important epigenetic modification, affecting a wide range of diseases and other

phenotypes [1-7], our finding that genetic or environmental interactions likely affect most mSNPs - and thus may also explain a substantial portion of the population specificity of DNA methylation levels, and their heritabilities - underscores the complex interplay of factors that influence epigenetic modifications. Further characterization of these factors will be critical for our understanding of the epigenome.

Materials and methods

Genome-wide DNA methylation analysis

Genomic DNA was purchased from the Coriell Institute. DNA concentration and purity were assessed spectrophotometrically using a NanoDrop ND-1000 (Thermo Scientific, Waltham, MA, USA). After random ordering of all samples, 1 μg of genomic DNA from each sample was bisulfite-converted using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA, USA) as per Illumina's Infinium specific protocol. Bisulfite converted DNA was then quantified by NanoDrop and concentrated to higher than 50 ng/μl using a Speedvac.

Quantitative DNA methylation measurements of bisulfite-treated genomic DNA were performed with the Infinium HumanMethylation27 BeadChip assay (Illumina, San Diego, CA, USA), using experimental procedures recommended by the manufacturer. Briefly, 200 ng of bisulfite-converted DNA was whole-genome amplified, fragmented by an enzymatic process and hybridized to BeadChip arrays. Two oligonucleotide probes interrogated each CpG site, one probe with sequences targeting methylated DNA and the other containing sequences targeting unmethylated DNA. After extension with DNP-labeled and biotin-labeled dNTP, each array was stained with Cy5 labeled anti-DNP antibodies and Cy3 labeled streptavidin and scanned with the Illumina iScan on a two-color channel to detect Cy3 labeled probes on the green channel and Cy5 labeled probes on the red channel. Using the Illumina GenomeStudio software package, methylation levels (β values) were then calculated by dividing the methylated probe signal intensity by the sum of methylated and unmethylated probe signal intensities. β values range from 0 (completely unmethylated) to 1 (fully methylated) and provide a quantitative readout of relative DNA methylation for each CpG site within the cell population being interrogated. This method was highly reproducible, as technical replicates across different runs had $r > 0.996$. All samples passed internal controls included on the HumanMethylation27 arrays, including controls for array background, hybridization quality, target specificity and bisulfite conversion. Furthermore, all samples passed our quality control check of having fewer than 5% of sites with either detection P-value < 0.05 or fewer than five beads being present on the array for a particular CpG site. Cluster analysis also indicated the absence of any outlier samples. Raw data have been deposited in the Gene Expression Omnibus database under accession number [GSE27146].

Samples from both populations were run together in a randomized order to avoid confounding batch effects with population differences. In order to test for the presence of batch effects, we tested whether the DNA methylation profiles of samples run in either the same batch number (1 to 4) or well number (1 to 96) were more similar to each other than expected by chance. Neither batch number nor well number was predictive of profile similarity (comparing correlation coefficients within batches or wells to all sample correlations, Wilcoxon $P = 0.79$ and 0.64, respectively), indicating the lack of any detectable batch effects.

Several steps were applied for normalization of β values across the subjects. First, average background intensity, as measured by negative background probes present on the array, was subtracted from the raw intensities to adjust for varying background signals across different samples. This background adjustment was done separately for raw data from the green and red channels to adjust for Cy3 and Cy5 differences. All negative intensities were assigned values of zero before further normalizations were performed. To minimize batch effects across different sets of arrays, background adjusted raw data from both channels were quantile normalized separately. Applying the same formula used by GenomeStudio, average β values were then recalculated using background subtracted and quantile normalized intensities of methylated probes divided by the sum of normalized intensities from unmethylated and methylated probes.

Pyrosequencing

DNA methylation of the promoter regions of *PLSCR2* and *IGSF2* containing specific CpG loci under the control of mSNPs were confirmed using bisulfite pyrosequencing. Genomic DNA (750 ng) was bisulfite converted using an EZ DNA Methylation Gold kit (Zymo Research). After PCR amplification of approximately 200 bp regions encompassing the target loci using specifically designed primers to ensure unbiased amplification, quantitative measurement of DNA methylation at each CpG was performed using a pyrosequencing primer located within 30 bp of the CpG interrogated. Reactions were measured on a PyroMark Q96 MD Pyrosequencer following the manufacturer's protocol, and analyzed using the Pyro Q-CpG software (Biotage, Uppsala, Sweden), which allows quality assessment of each measurement. CpG loci that were called 'passed' in the default software settings are shown in Figure 1b ($n = 175$ for *IGSF2*; $n = 156$ for *PLSCR2*). To

assess the agreement between methods, we used Pearson's correlation (as throughout the manuscript), because rank-based correlations do not account for the clustering of most samples within a small range of methylation (for example, 95 to 100% methylation for *PLSCR2* in Figure 1b). An alternative metric, classifying sites into high or low methylation based on a cutoff and measuring agreement in a 2 × 2 contingency table, led to results similar to the Pearson correlation across a wide range of cutoffs (data not shown). Primer sequences used for DNA amplification and pyrosequencing are available upon request.

Calculation of false discovery rates

All FDRs were estimated by randomization, which preserves all aspects of the data that might affect statistical analyses. For example, the FDR for population-specific methylation was estimated by randomly assigning CEU/YRI labels, and recalculating the Wilcoxon *P*-value on the randomized data (resulting in an essentially uniform distribution of *P*-values, like that shown in Figure 1c). FDRs for mSNPs were estimated by pairing genotypes with randomly chosen methylation profiles, and calculating mSNPs as for the real data. Because of the family trio structure of the HapMap samples, not all samples are independent; to account for this in our randomization procedure, we also performed randomizations based on swapping methylation data for entire trios, in effect treating each trio as an independent unit composed of three methylation profiles and three genome sequences. This procedure yielded indistinguishable FDRs compared to randomizing all samples individually. All FDRs are based on at least 1, 000 randomizations.

Heritability analysis

Narrow-sense heritabilities (h^2) were estimated as the correlation between average parental values and their offspring. Because the offspring and parental variances are equal, this is equivalent to performing regression. Although heritabilities are by definition non-negative, our estimates are often negative due to the limited power inherent in our data. We note that our method of estimating h^2 assumes that there is no shared environmental variance between parents and offspring that impacts DNA methylation; if this assumption is violated, we will overestimate h^2 (with an upper bound of H^2, the broad-sense heritability). It also assumes that somatic DNA methylation is not passed directly from parent to offspring through the germline, since this would violate the assumptions of the heritability estimation. To estimate the number of CpG sites with heritable methylation, we generated 1, 000 randomized versions of the h^2 distribution (see above), and calculated the number of sites with greater methylation in the real data, compared to each randomized distribution. Visually, this corresponds to the area in between the two distributions, on the right side (positive values) where the real distribution is shifted to the right. The average difference across the 1, 000 randomizations was 762 sites for CEU, and 930 for YRI. Note that this procedure allows us to estimate the number of heritable sites, but not specify which specific sites are the heritable ones; thus, it is not possible to calculate an FDR for these estimates.

mSNP analysis

mSNPs were identified by calculating correlations between SNP genotypes (arbitrarily coded as 0, 1, and 2) and methylation levels. Only SNPs within 100 kb of each CpG site were tested, to reduce the multiple testing burden. Although the 1000 Genomes SNP catalog is more complete, we used HapMap genotypes [16] for the mSNP analysis, since not all cell lines for which we collected methylation data have been sequenced as part of the 1000 Genomes Project [22]. We required a minimum of 5 minor alleles among the 90 individuals of each population to include a SNP in this analysis (for details of how we accounted for the family trio structure, see 'Calculation of false discovery rates' above). This resulted in 2, 668, 982 YRI SNPs and 2, 405, 735 CEU SNPs (1, 969, 973 shared by both). For the analysis of genetic variants contributing to population-level differences, only the SNPs shared by both populations were used, and population was represented in the multiple regression as 0/1 for CEU/YRI.

Correlations were recorded as the absolute value of the correlation coefficient, since the sign is arbitrary, depending on how genotypes are coded as 0/1/2. However, for comparisons between CEU and YRI correlations, the fact that all correlations are positive means that the difference between associations can be underestimated. If the same SNP (or two SNPs in high LD) was used to calculate the correlation with a particular CpG site's methylation in both populations, the signs could be used; however, in most cases a site's strongest correlation was with different SNPs in CEU and YRI, precluding the use of signs.

Bisulfite sequencing of *RNF186* promoter region

Genomic DNA (500 ng) was bisulfite converted using the EZ-96 DNA Methylation Gold Kit (Zymo Research) as per the manufacturer's protocol with minor modifications. A 532 bp region upstream of the *RNF186* gene containing the SNP rs3806308 and the CpG site cg09195271 from the IlluminaHuman Methylation array was amplified by nested PCR reactions using Hotstar Taq (Qiagen, Hilden, Germany). The first round of PCR amplification was done using 55°C annealing temperature for 30 cycles and the primer pair F3

(GGATATAGAGGGTGGTTTGTAGTGTTAGT) and R2 (ACRCACAAATATTTAACACCTACTACT). A 3 μl aliquot of the material obtained in the first round was further amplified in the second round in a total volume of 50 μl, using 51°C annealing temperature for 35 cycles and the primer pair F2 (TGAATGAAATATTTGTTTGAGGGAGTGT) and R3 (CCTTAAAACCACAACTATTATATTCACAA). All primers were designed to be specific for bisulfite converted DNA. The amplified PCR product was separated from primers by electrophoresis in a 1.5% Tris-acetate-EDTA (TAE) agarose gel, excised and purified using the QIAquick gel extraction kit (Qiagen). Purified DNA was then ligated into plasmid pGem-T Easy using the pGem-T Easy vectory system (Promega, Madison, WI, USA) and transformed into competent JM109 *Escherichia coli* (Promega) by the $CaCl_2$ method. Colonies carrying a plasmid containing an insert were then selected based on blue-white screening. Plasmid DNA was extracted using Qiaprep Spin Miniprep kit (Qiagen). Plasmid clones containing the appropriate sized insert, as determined by a restriction digestion analysis, were sequenced using T7 and/or SP6 primers by Genewiz Inc. South Plainfield, NJ, USA. Sequences were analyzed using Sequencher sequence analysis package 4.6 (Gene Codes Corporation, Ann Arbor, MI, USA).

Additional material

Additional file 1: Supplemental text, Tables S1 and S2, and Figures S1 to S19 [27-30].

Abbreviations

bp: base pair; CEU: HapMap population of Northern European ancestry; CpG: cytosine-phosphate-guanine; FDR: false discovery rate; GxE: gene-by-environment; GxG: gene-by-gene; LCL: lymphoblastoid cell line; LD: linkage disequilibrium; mSNP: methylation-associated SNP; SNP: single-nucleotide polymorphism; YRI: HapMap population of Yoruban ancestry.

Acknowledgements

We thank M Feldman, M Hayden, J Rine, S Roy, and an anonymous reviewer for helpful comments and discussion. We further thank M Lorincz for advice on bisulfite sequencing and use of Sequencher software, and A Devlin for use of the PyroMarkMD system. Work in MSK's laboratory is supported by National Institute of Health (NIH) grant R24MH-081797-01. Work in HBF's laboratory is supported by National Institute of Health (NIH) grant 1R21HG005750-01A1. MSK is a Scholar of the Canadian Institute for Advanced Research and of the Mowafaghian Foundation. HBF is an Alfred P Sloan Fellow and Pew Scholar in the Biomedical Sciences.

Author details

[1]Department of Biology, Stanford University, Stanford, CA 94305, USA. [2]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, V6T 1Z3, Canada. [3]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Vancouver, British Columbia V5Z 4H4, Canada.

Authors' contributions

MSK designed the project, oversaw data generation and wrote the paper. HBF designed the project, analyzed the data and wrote the paper. LL and SN generated and normalized the data. All authors have approved the final manuscript for publication.

Competing interests

The authors declare that they have no competing interests.

Received: 29 September 2011 Revised: 30 January 2012
Accepted: 9 February 2012 Published: 9 February 2012

References

1. Mohn F, Schübeler D: **Genetics and epigenetics: stability and plasticity during cellular differentiation.** *Trends Genet* 2009, **25**:129-136.
2. Bonasio R, Tu S, Reinberg D: **Molecular signals of epigenetic states.** *Science* 2010, **330**:612-616.
3. Law JA, Jacobsen SE: **Establishing, maintaining and modifying DNA methylation patterns in plants and animals.** *Nat Rev Genet* 2010, **11**:204-220.
4. Illingworth RS, Bird AP: **CpG islands - 'a rough guide'.** *FEBS Lett* 2009, **583**:1713.
5. Chang SC, Tucker T, Thorogood NP, Brown CJ: **Mechanisms of X-chromosome inactivation.** *Front Biosci* 2006, **11**:852-866.
6. Jaenisch R, Bird A: **Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.** *Nat Genet* 2003, **33(Suppl)**:245-254.
7. Bjornsson HT, Fallin MD, Feinberg AP: **An integrated epigenetic and genetic approach to common human disease.** *Trends Genet* 2004, **20**:350-358.
8. Zilberman D, Henikoff S: **Genome-wide analysis of DNA methylation patterns.** *Development* 2007, **134**:3959-3965.
9. Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, Zheng H, Yu J, Wu H, Sun J, Zhang H, Chen Q, Luo R, Chen M, He Y, Jin X, Zhang Q, Yu C, Zhou G, Sun J, Huang Y, Zheng H, Cao H, Zhou X, Guo S, Hu X, Li X, Kristiansen K, Bolund L, Xu J, *et al*: **The DNA methylome of human peripheral blood mononuclear cells.** *PLoS Biol* 2010, **8**:e1000533.
10. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S: **DNA methylation profiling of human chromosomes 6, 20 and 22.** *Nat Genet* 2006, **38**:1378-1385.
11. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, Craig DW, Redman M, Gershon ES, Liu C: **Genetic control of individual differences in gene-specific methylation in human brain.** *Am J Hum Genet* 2010, **86**:411-419.
12. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, Johnson R, Zielke HR, Ferrucci L, Longo DL, Cookson MR, Singleton AB: **Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain.** *PLoS Genet* 2010, **6**:e1000952.
13. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK: **DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.** *Genome Biol* 2011, **12**:R10.
14. Boks MP, Derks EM, Weisenberger DJ, Strengman E, Janson E, Sommer IE, Kahn RS, Ophoff RA: **The relationship of DNA methylation with age, gender and genotype in twins and healthy controls.** *PLoS One* 2009, **4**:e6767.
15. Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, McRae AF, Visscher PM, Montgomery GW, Gottesman II, Martin NG, Petronis A: **DNA methylation profiles in monozygotic and dizygotic twins.** *Nat Genet* 2009, **41**:240-245.
16. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, *et al*:

Integrating common and rare genetic variation in diverse human populations. *Nature* 2010, **467**:52.
17. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG: **Common genetic variants account for differences in gene expression among ethnic groups.** *Nat Genet* 2007, **39**:226-231.
18. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM: **Gene-expression variation within and among human populations.** *Am J Hum Genet* 2007, **80**:502-509.
19. Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME: **Evaluation of genetic variation contributing to differences in gene expression between populations.** *Am J Hum Genet* 2008, **82**:631-640.
20. Grafodatskaya D, Choufani S, Ferreira JC, Butcher DT, Lou Y, Zhao C, Scherer SW, Weksberg R: **EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines.** *Genomics* 2010, **95**:73-83.
21. Caliskan M, Cusanovich DA, Ober C, Gilad Y: **The effects of EBV transformation on gene expression levels and methylation profiles.** *Hum Mol Genet* 2011, **20**:1643-1652.
22. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467**:1061.
23. Lynch M, Walsh B: **Genetics and Analysis of Quantitative Traits.** *Sinauer Assoc* 1997.
24. Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC: **Sequence polymorphisms cause many false cis eQTLs.** *PLoS One* 2007, **2**:e622.
25. Akey JM, Biswas S, Leek JT, Storey JD: **On the design and analysis of gene expression studies in human populations.** *Nat Genet* 2007, **39**:807-808.
26. Friso S, Choi SW: **Gene-nutrient interactions in one-carbon metabolism.** *Curr Drug Metab* 2005, **6**:37-46.
27. Bjornsson HT, Sigurdsson MI, Fallin MD, Irizarry RA, Aspelund T, Cui H, Yu W, Rongione MA, Ekström TJ, Harris TB, Launer LJ, Eiriksdottir G, Leppert MF, Sapienza C, Gudnason V, Feinberg AP: **Intra-individual change over time in DNA methylation with familial clustering.** *JAMA* 2008, **299**:2877-2883.
28. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464**:768.
29. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK: **High-resolution mapping of expression-QTLs yields insight into human gene regulation.** *PLoS Genet* 2008, **4**:e1000214.
30. Chen YA, Choufani S, Ferreira JC, Grafodatskaya D, Butcher DT, Weksberg R: **Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray.** *Genomics* 2011, **97**:214-222.

doi:10.1186/gb-2012-13-2-r8
Cite this article as: Fraser *et al.*: Population-specificity of human DNA methylation. *Genome Biology* 2012 **13**:R8.