

# The heritability and patterns of DNA methylation in normal human colorectum



HMG Advance Access published March 2, 2016



# The heritability and patterns of DNA methylation in normal human colorectum

**Amy Rowlatt[1], Gustavo Hernandez-Sanchez[2], Maria Carolina Sanabria[3,4], Martha Serrano-Lopez[3,4], Konrad Rawlik[1], Eva Hernandez-Illan[5], Cristina Alenda[6], Adela Castillejo[5], Jose Luis Soto[5], Chris S. Haley[7], Albert Tenesa[1,7,*]**

[1]Roslin Institute at the University of Edinburgh, Edinburgh, Midlothian, UK

[2]Grupo de Investigacion Epidemiologica, Instituto Nacional de Cancerologia de Colombia, Bogota, Colombia

[3]Grupo de Investigacion en Biologia del Cancer, Instituto Nacional de Cancerologia, Bogota, Colombia

[4]Departamento de Quimica, Universidad Nacional de Colombia, Bogota, Colombia

[5]Unidad de Investigación, Hospital General Universitario de Elche, 03203 Elche, Alicante, Spain

[6]Pathology, Alicante University Hospital, Alicante, Spain

[7]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, Midlothian, EH4 2XU, UK

[*]Correspondence: The Roslin Institute, The University of Edinburgh, Easter Bush, Roslin, Midlothian, EH25 9RG, Scotland. Tel: 0044 (0)131 651 9100, Fax: 0044 (0)131 651 9220, E-mail: Albert.Tenesa@ed.ac.uk

## Abstract

DNA methylation (DNAm) has been linked to changes in chromatin structure, gene expression and disease. DNAm level can be affected by genetic variation; although, how this differs by CpG dinucleotide density and genic location of the DNAm site is not well understood. Moreover, the effect of disease causing variants on DNAm level in a tissue relevant to disease has yet to be fully elucidated. To this end, we investigated the phenotypic profiles, genetic effects and regional genomic heritability for 196080 DNAm sites in healthy colorectum tissue from 132 unrelated Colombian individuals. DNAm sites in regions of low CpG density were more variable, on average more methylated and were more likely to be significantly heritable when compared to DNAm sites in regions of high CpG density. DNAm sites located in intergenic regions had a higher mean DNAm level and were more likely to be heritable when compared to DNAm sites in the transcription start site (TSS) of a gene expressed in colon tissue. Within CpG dense regions, the propensity of DNAm level to be heritable was lower in the TSS of genes expressed in colon tissue than in the TSS of genes not expressed in colon tissue. In addition, regional genetic variation was associated with variation in local DNAm level no more frequently for DNAm sites within colorectal cancer (CRC) risk regions than it was for DNAm sites outside such regions. Overall, DNAm sites located in different genomic contexts exhibited distinguishable profiles and may have a different biological function.

## Introduction

Cytosine DNA methylation (DNAm) is a covalent modification of DNA brought about by the addition of a methyl group to the 5$^{th}$ position of the pyrimidine ring of cytosine. In differentiated mammalian cells DNAm occurs almost exclusively at cytosine bases directly upstream of guanine bases (CpG dinucleotide) (1). Importantly, DNAm can be mitotically stable. During embryogenesis the epigenome is erased and reprogrammed, initially prior to blastocyst formation and subsequently in the germ cells (Reviewed in (2-4)). In the somatic cells of the developed organism Dnmt1 is the main methyltransferase which targets hemi-methylated DNAm during replication (5). Dnmt1 interacts with a host of proteins at the replication foci including histone deacetylases (6) and histone methyltransferases (7, 8) signifying the complex relationship between modification and organization of the chromatin and maintenance of DNAm level throughout cell division.

Studies have shown that DNAm level is linked to gene expression level (9, 10). At promoter regions, within a population of individuals average DNAm level and average gene expression level of an associated gene were negatively correlated across genes (9). However, at any given DNAm site, even at a DNAm site within a promoter region, there may be a positive or a negative association between that DNAm site and expression levels of an associated gene across the population of individuals (9). Additionally, DNAm level has been found to associate with sex and age (10-12) and several environmental factors including early life socioeconomic status and stress (10). Differences in DNAm levels have been observed for DNAm sites in different functional genomic contexts. For instance, DNAm levels tend to decrease towards the TSS of a gene and are relatively high throughout the gene body (1, 13, 14). Moreover, changes in DNAm levels have been found between DNAm sites located in CpG dense regions of the genome (CpG Islands) and those DNAm sites located outside CpG Islands, whether they are close to (CpG Island Shores and CpG Island Shelves) or distant from such islands (13). Furthermore, results from genome-wide association studies, twin studies and studies utilizing mixed linear models indicate that the genotype can affect level of DNAm (15-22) and that *cis* acting genetic variation can explain a substantial proportion of the phenotypic variation for some

DNAm sites (16, 17, 19, 21, 22). These studies have been conducted on a limited number of easy to access tissues and report a wide range of heritability estimates for site-specific DNAm level across the genome.

The average heritability estimate of site-specific DNAm level varies across tissues and it is likely to be influenced by the method used for estimation and the estimates should be interpreted accordingly. For example, monozygotic (MZ) and dizygotic (DZ) twin studies of cells from buccal epithelia and white blood cells estimated the average heritability ($\hat{h}^2_{ped}$) of all assayed site-specific DNAm levels to be 0.30 and 0.01, respectively(15) Extended families provided an estimate of average $\hat{h}^2_{ped}$ for DNAm levels in peripheral blood lymphocytes of 0.20 (20) . These studies capture the full extent of the additive genetic heritability. However, sources of shared environmental variance among family members could bias estimates of the additive genetic variance when the environmental variance is un-modelled or difficult to disentangle. The extent of the common environmental effect was tested for peripheral blood lymphocytes (20). The genomic heritability ($\dot{h}^2_g$), the proportion of the phenotypic variation that can be explained by genetic variation measured by tagging single nucleotide polymorphism (SNP) on a genotyping array can be estimated using mixed linear models and nominally unrelated individuals (23). However, insufficient linkage disequilibrium between a SNP and the causal variation can mean that not all causal variation, particularly rare causal variation, may be captured by this method. Therefore, $\dot{h}^2_g$ may be biased downwards in relation to the true additive heritability. When using this method it is straightforward to partition the genetic variation into genomic regions (regional heritability, $\dot{h}^2_{g,r}$) by simultaneously modelling the effects of SNP within a region of interest (24). Advantages of this method over independently testing the association of each SNP in the genome with a trait (SNP by SNP genome-wide association study) are that small effects within a region maybe compounded into a measureable estimate and effects of rare variants present on a particular haplotype may be captured more effectively (24). Indeed, simulations have shown that in some instances the estimator $\dot{h}^2_{g,r}$ is more accurate than estimates of heritability obtained from a SNP

by SNP genome-wide association study (25). A recent study investigated the effect of the size of the genomic region surrounding a DNAm site on the estimator $\ddot{h}^2_{g,r}$ for DNAm levels measured in the cerebellum, frontal cortex, pons and the temporal cortex for 150 unrelated individuals (22). The authors tested seven region sizes: 10kb, 50kb, 100kb, 500kb, 1MB, the local chromosome and the whole genome, and found that using a region size of 50kb centered around the DNAm site produced the greatest number of significant $\ddot{h}^2_{g,r}$ (22). For instance, depending on the tissue, using the local chromosome or the whole genome reduced the number of significant results to 100 or less whereas using a region of 50kb produced between 600 and 812 significant results (22). The increase in power that accompanies the optimal region size comes from the smaller number of SNPs modelled within the target region compared to the whole genome. Regions too small will not capture the causative variation. However, because a substantial proportion of the causative genetic variation for DNAm level is thought to act in *cis*, regions too large will include extraneous SNPs that add noise to the estimator, $\ddot{h}^2_{g,r}$ (22). The average site-specific $\ddot{h}^2_{g,r}$ for regions of +/- 50KB around the DNAm site and for DNAm sites with a significant $\ddot{h}^2_{g,r}$ is 0.30 for DNAm levels measured in the cerebellum, frontal cortex, pons and the temporal cortex (22).

The extent to which genetic variation affects variation in DNAm level may also differ within tissue depending on the functional genomic context of the DNAm sites. For instance, the differences in DNAm levels that have been observed for DNAm sites in different genomic contexts could relate to a difference in the control of DNAm level. A recent study assessed $\hat{h}^2_{ped}$ of site-specific DNAm for DNAm sites located in high and low CpG density regions of the genome in peripheral blood lymphocytes (20). This study revealed that estimates in regions of high CpG density were 0.127 or 0.158, whereas in regions of low CpG density estimates were greater, 0.235 or 0.223, depending on which probe type (Infinium I or Infinium II) was used to assay the DNAm level. In human brain tissue, the estimator, $\hat{h}_{r,g}{}^2$ showed an increased proportion of heritable DNAm sites in regions of the genome with low CpG density compared to high CpG density (22). In addition, this same study found a decreased proportion of heritable DNAm sites local to genes upregulated in a tissue specific manner

compared to genes expressed ubiquitously across tissues(22). While these studies have begun to explore the extent of heritability for DNAm sites located in different genomic contexts they have been conducted in a minority of tissues and have considered a limited selection of functional subgroups for DNAm sites.

To build on the work of others who have investigated the phenotypic differences and control of DNAm in different genomic contexts, we assayed 196080 DNAm sites with the Infinium HumanMethylation450K BeadChip (HM450K) in healthy colorectum tissue collected from 132 unrelated Colombian subjects who attended Colonoscopy examination and with diagnosis of hyperplasic polyp, adenoma, in situ carcinoma or carcinoma of the rectum or colon. A whole cell biopsy was taken from the healthy colonic tissue from one of the following locations: ascending, transverse, descending, or sigmoid colon, cecum, rectum or region where the sigmoid colon joins the rectum. We grouped the DNAm sites based on location in relation to CpG density, expression status and functional regions of genes. We refer to these groups collectively as contextual groups. Within each contextual group we assessed the profile of mean site-specific DNAm level where mean site-specific DNAm level refers to the average DNAm level for a given DNAm site calculated across all the 132 samples. Subsequently we estimated the effect of local genetic variation on site-specific DNAm level, using a region size of +/- 1MB surrounding the DNAm site following earlier work of others (16, 21). We used a regional heritability approach (24) and estimated the proportion of variability in site-specific DNAm levels that is due to local genetic variation, $\hat{h}_{r,g}^2$ . We also contrasted the distribution of $\hat{h}_{r,g}^2$ for DNAm sites within and outwith known susceptibility loci for Colorectal Cancer (CRC, OMIM #114500). In addition, CRC can manifest in colorectal epithelial cells (26), which are significantly more costly and challenging to extract from the colon than whole cell biopsies. To establish the extent of the difference in the regulation of DNAm levels in colon epithelial tissue and whole cell biopsy, we studied the phenotypic profiles and $\hat{h}_{r,g}^2$ for DNAm level of genes expressed in epithelial cells obtained from laser capture microdissection (LCM) and whole colon biopsy (WCB).

## Results

### Average DNAm Level and relationship to CpG Density, Genic Location and Gene Expression

In the manifest file for the HM450K array each DNAm site is annotated as being located either within a CpG Island (island), within 2kb upstream or downstream of a island (north shore and south shore respectively), within 2-4kb upstream or downstream of an island (north shelf and south shelf respectively) or none of the aforementioned categories which we term sea. An island was defined as being composed of one or more adjacent sections of the genome each 500bp in length with a C and G density greater than 50% and an observed to expected ratio of CpG dinucleotides greater than 0.60 (27). We grouped our 196080 DNAm sites based on aforementioned HM450K array annotation (Table 1). Using DNAm level values adjusted for gender, age, batch, diagnosis, localization, and two genotype principal components we found a substantial difference in the distribution of average site-specific DNAm level across contextual groups of varying CpG density (Figure 1). The average site-specific DNAm level of DNAm sites in islands tended to be much lower than that of DNAm sites located in the sea (mean and median M-value was -2.67 and -3.51, and 1.50 and 1.89 for islands and sea, respectively). Additionally, our results showed that the distribution of average DNAm level for DNAm sites in the north and south shores were similar to one another and more similar to the distribution of average DNAm level for DNAm sites in islands rather than DNAm sites in the sea (Figure 1). Conversely, the distribution of average DNAm level for DNAm sites located in the north and south shelves were similar to one another and were more similar to the distribution observed for DNAm sites located within the sea rather than within islands (Figure 1). Additionally, we found that within shores the mean site-specific DNAm level is a function of the distance from the edge of the island. Mean site-specific DNAm level increased with distance from the edge of the island in a non-linear fashion (Figure 2). However, this relationship is not observed for DNAm sites located within shelves (Figure 2).

Unless otherwise specified, the following analyses were based on the two most extreme cases of CpG density: high CpG density regions (islands) and low CpG density regions (sea). We tested if the DNAm level of DNAm sites located in the transcription start site (TSS) of genes expressed in WCB was different to those located in the TSS of genes expressed in cells from the colon epithelium collected using LCM. The genes that were expressed in the LCM and the WCB were excluded from the WCB group for this analysis. We found that DNAm sites in the TSS of genes expressed in WCB had an M-value that was on average 0.15 greater than DNAm sites located in the TSS of genes expressed in LCM (mean WCB = -3.32, mean LCM = -3.47, T-test P= $2.41*10^{-7}$).

Subsequently, we grouped DNAm sites located within the sea or an island into four mutually exclusive sets based on location a) in a transcription start site (TSS) of a gene expressed in WCB b) in a TSS of a gene that is not expressed in WCB c) in intragenic DNA, where we do not distinguish between genes expressed or not expressed in colon because the methylation level of intragenic DNAm sites has not been correlated with the expression of the surrounding gene or d) intergenic DNA (

Table 2). Additionally, we choose to use the full set of genes expressed and not expressed in WCB rather than exclusively in colonic epithelial cells because DNAm level was assayed from WCB. We refer to each of the eight contextual groups individually as: island TSS expressed (within an island and a TSS of a gene expressed in colon), island TSS not expressed (within an island and in a TSS of a gene not expressed in colon), island intragenic (within an island and intragenic), island intergenic (within an island and intergenic), sea TSS expressed (within the sea and the TSS of a gene expressed in colon), sea TSS not expressed (within the sea and the TSS of a gene not expressed in colon), sea intragenic (within the sea and intragenic) and sea intergenic (within the sea and intergenic). Within each of the eight contextual groups we investigated the distribution of mean site-specific DNAm level.

We found a significant difference in the distribution of mean site-specific DNAm level between island TSS expressed and island TSS not expressed (Kolmogorov-Smirnov test; $P<2.16*10^{-16}$) and between sea TSS expressed and sea TSS not expressed (Kolmogorov-Smirnov test; $P<2.16*10^{-16}$) (Figure 3). We compared the mean of the sea TSS expressed to that of the sea TSS not expressed and we compared the mean of the island TSS expressed to that of island TSS not expressed. These two comparisons were both statistically significant (T-test, $P<2.16*10^{-16}$, $P<2.16*10^{-16}$) and in both cases being located in the TSS of genes not expressed in colon led to an overall greater mean site-specific DNAm level. Additionally, we found that the mean of the distribution of mean site-specific DNAm level for DNAm sites located in intragenic and intergenic regions was greater than for DNAm sites located in a TSS of a gene (Figure 4).

We explored the variation in DNAm level by investigating the extent to which residual site-specific DNAm level varied across the 132 samples for the full set of 196080 DNAm sites that passed quality control procedure. The variance of residual DNAm level ranged between $3.00x10^{-3}$ and 4.91. We observed that on average over all sites the variance of DNAm level was significantly higher for DNAm sites located within the sea than for DNAm sites located within an island (Mean for sea and island respectively: 0.198 and 0.152, $P < 2.2*10^{-6}$ Figure 5). Moreover, we found the following

pattern for the magnitude of mean residual variance both within islands and sea: TSS Expressed < TSS Not Expressed < Intragenic < Intergenic. Within the islands the difference in mean residual variance was highly significant ($P < 1.00*10^{-8}$) for all pairwise comparisons of these categories. Within the sea the difference in mean residual variance was at least moderately significant ($P < 0.001$) for all pairwise comparisons made except for the comparison of intragenic and intergenic. In this case, the mean residual variance for the sea intragenic and sea intergenic was not significantly different ($P=0.240$).

**Heritabilities of Site-specific DNAm Level**

We found that at 20239 DNAm sites, 10.32% of the 196080 tested, SNPs within 1MB explained a significant proportion of the variation in methylation level (nominal $P < 0.05$). The percentage of heritable DNAm sites exceeds that expected from a false positive rate of 0.05 under the null hypothesis that DNAm level is not associated with local genetic variation. For significantly heritable loci, the proportion of the variance in DNAm under local genetic control ranged between 0.06 and 0.99 with a mean of 0.29 and median of 0.26 (Figure 6). We found that the numbers of SNPs in the local region (Figure 7) explain a minute but significant proportion of the variance in heritability estimates for the DNAm sites with a significant heritability (Univariate Linear Regression: $R^2 = 0.005$, slope = $2.372*10^{-5}$, $P < 2.2*10^{-16}$). For instance, considering the range of the number of SNPs within a region, at the first decile (304 SNPs) and ninth decile (2733 SNPs) we expect a respective $7.2*10^{-3}$ and $6.5*10^{-2}$ increase in the heritability for the DNAm sites found to be significantly associated with local genetic variation. In addition, we found that the variance of residual DNAm level at each DNAm site explained a proportion of $1.93*10^{-2}$ of the variance in the heritability estimate (Univariate Linear Regression: $P < 2.2*10^{-16}$). An ANOVA indicated that genomic context explained a significant proportion ($P < 2.2*10^{-16}$) of the variability in the heritability estimate after accounting for residual variance.

**Heritability of DNAm in genes expressed in Whole Colorectal Biopsies and Colorectal Epithelial**

We investigated the heritability of site-specific DNAm level for genes expressed in LCM and in WCB excluding those expressed in LCM (Figure 8). The difference between the average mean site-specific DNAm level for the significantly heritable DNAm sites within the two groups was not significant (mean LCM = 0.272, mean WCB = 0.283, T-test P =P=0.151). Additionally, the proportion of significantly heritable DNAm sites in the LCM and WCB group was 0.0723 and 0.0734 respectively and was not significantly different from one another (P = 0.0834).

**Heritability of DNAm Sites in Whole Colorectal Biopsies by Genomic Context**

The proportion of sites with a significant heritability was higher in the sea than in islands ($P < 2.20*10^{-16}$, Table 3). This result was driven by the difference between DNAm sites located within the TSS of a gene or in intragenic regions. The proportion of heritable sites is 1.54 times higher for DNAm located in the TSS of the sea than in the TSS of an island ($P < 2.2*10^{-16}$); additionally, the proportion of heritable DNAm sites is 1.13 times higher for DNAm sites located in the sea intragenic than the island intragenic contextual group ($P = 1.47*10^{-6}$). There was no significant difference in the proportion of heritable DNAm sites located in intergenic regions when comparing between the sea and island contextual groups (P = 0.125).

The proportion of heritable DNAm sites was lower for sea intragenic than the other three sea contextual groups. This difference was highly significant for the comparison of sea intragenic and sea intergenic ($P = 1.55*10^{-14}$) and for the comparison of sea intragenic and sea TSS not expressed ($P = 2.65*10^{-5}$). The difference was significant at a nominal threshold ($P < 0.05$) for the sea intragenic and sea TSS expressed comparison ($P=1.52*10^{-2}$). The proportion of heritable DNAm sites was significantly different for all comparisons made within the island contextual groups. This difference was highly significant ($P < 1.00*10^{-8}$) except for the island TSS not expressed and island intragenic comparison ($P = 4.34*10^{-2}$). Within the island contextual groups the proportion of heritable DNAm sites was as follows: intergenic > intragenic > TSS Not expressed > TSS Expressed.

The proportion of heritable DNAm sites for each contextual group followed a similar pattern to the mean local genetic variance for each contextual group (Figure 9). Across each four sea and island contextual groups the proportion of heritable DNAm sites was correlated with the mean local genetic variance (Pearson's correlation: r = 0.700 and r = 0.999 for sea and island respectively).

The average $\hat{h}^2_{r,g}$ for significantly heritable DNAm sites located within each of the genomic contexts was similar (

Table 3). DNAm sites significantly associated with local genomic variation and located within an island were on average 0.9% less heritable than DNAm sites located in the sea.

**SNP by SNP GWAS for DNAm level at 196080 DNAm sites**

We conducted a GWAS for each of the 196080 DNAm sites to localize causal variation within each of the *cis* regions and to determine if there was evidence for genetic effects on DNAm level in *trans*. We found enrichment for *cis* and *trans* genetic effects on DNAm level beyond what would be expected by chance assuming the test for association of each SNP and DNAm site are independent (

Table 4). DNAm sites with at least one significant *cis* SNP association were found on average to have a minimum of two associated *cis* SNPs, where the average number depended on the threshold specified (

Table 4). Each DNAm site with a significant RH estimate ($P < 0.05$, n=20239) was paired with the SNP to which it was most significantly associated with in *cis*. We calculated the proportion of the regional heritability estimate explained by variation of the SNP for each pair (Figure 10). In, the majority of cases (95.8%) the region explained equal or more variance in the DNAm level than the single most significant SNP (Figure 10).

**Heritability of DNAm Sites in Whole Colorectal Biopsies with Respect to Loci Associated with Colorectal Cancer**

An extensive number of genetic variants have been found to associate with complex disease including CRC (28). However, in the majority of cases how the identified genetic risk variants act to increase disease susceptibility is unknown. Genetic variation could increase risk to disease by mediating changes in DNAm level in healthy tissue. If an association between a susceptibility SNP and DNAm level in healthy tissue is obtained, one possibility is that the variation in DNAm level interacts with additional variables, such as environmental factors, to subsequently lead to disease. Therefore, we sought to determine if variation in DNAm level in healthy colon tissue was associated with genetic variation that incurs susceptibility to CRC. To this end a total of 83 unique autosomal SNPs identified as associating with CRC were downloaded from the NHGRI GWAS catalogue (28). We defined a region of +/- 1MB surrounding each SNP associated with CRC as a risk region. A total of 10469 DNAm sites were located within a defined risk region and due to our definition of a risk region the calculation of $\ddot{h}^2_{r,g}$ included the effects at the position of the risk SNP. Indeed, we found that an equal proportion (0.103) of DNAm sites were heritable within and outwith risk regions. The average heritability was also similar for DNAm sites located within and outwith risk regions (within = 0.290, outwith = 0.292).

In conjunction, a recent study (29) found that DNAm levels of two DNAm sites measured in healthy colorectal tissue, cg15193198 and cg24112000, were associated (FDR < 0.05) with the local CRC risk variant rs4925386 located on chromosome 20 at 60921044 base pairs. In our study, both cg15193198

and cg24112000 were significantly heritable both when including rs4925386 in the calculation of the genetic relationships and when excluding rs4925386 (

Table 5). We determined that rs4925386 explained 92% and 55% of the $\ddot{h}^2_{g,r}$ estimate for cg15193198 and cg24112000 respectively and $\ddot{h}^2_{g,r}$ was still significant for cg24112000 when fitting rs4925386 as a fixed effect (

Table 5)

## Discussion

DNAm sites located in different genomic contexts with respect to CpG density and genic location exhibit unique profiles in the human colorectum. We have shown that average DNAm level is related to CpG density and genic location. The relationship of DNAm level with CpG density has been observed for DNAm level measured at promoters in peripheral blood mononuclear cells and fibroblasts (10, 30) and concurs with that found by a recent study profiling DNAm level in 17 somatic tissues (13). In addition, we find that specifically within shores that there is a shift in average DNAm level from predominantly unmethylated to methylated as distance increases from the edge of the CpG dense islands. This change in DNAm level is suggestive of a transitional zone at the edge of islands captured by the definition of shore. Overall, the lower and less variable average site-specific DNAm level of DNAm sites located in islands compared to sea is consistent with the traditional view that CpG dense regions of the genome persist due to low methylation and a reduced rate of spontaneous deamination and transition that is typically higher for methylated CpG dinucleotides (31). Our finding that irrespective of CpG density, DNAm level was lower in the TSS than within intergenic or intragenic regions concurs with what has been observed in H1 embyronic cells where DNAm level has been shown to decrease between the promoter and 5'UTR region before increasing through the gene body and into the 3'UTR (1). The level of DNAm has also been shown to be greater in intragenic and intergenic regions compared to promoter regions in human brain frontal cortex grey matter (32). Lower variation of DNAm level within TSS compared to with the intragenic and intergenic regions and within islands supports the idea that DNAm in CpG dense regions of the genome and in TSS target housekeeping genes (30, 33). Housekeeping genes are essential for normal cell maintenance and thus expression of these genes may be tightly regulated and this could be reflected by the low level and low variation of site-specific DNAm level in these regions. Additionally, DNAm sites located in the TSS of a gene not expressed in WCB were on average more methylated than DNAm sites located in the TSS of a gene that was expressed in WCB. This result is

suggestive of an overall inverse correlation between mean gene expression and mean DNAm level which has previously been observed (9).

We have assessed, on a genome-wide scale, the local heritability of site-specific DNAm level in normal WCB using unrelated Colombian individuals. A total of 10.32% of DNAm sites in WCB were significantly affected by local genetic variation. The mean $\hat{h}^2_{r,g}$ for the heritable sites was 0.29 but the estimates vary substantially with some DNAm sites exhibiting a low heritability and some DNAm sites exhibiting heritability close to one. The implication is that DNA methylation level can be inherited through the germ-line. These results are consistent with previous estimates of the number of DNAm sites and gene expression probes across the genome affected by local genetic variation and with the wide range of heritability estimates reported for levels of DNAm and gene expression (16, 22, 34-36). Indeed, we found that that there were an increased proportion of significantly heritable DNAm sites located in the sea compared to islands. Overall, our finding that heritable DNAm sites were enriched for location outside of islands is in accordance with what is observed in human brain tissue (22).We hypothesize that the substantial difference in the mean estimates of $\hat{h}^2_{ped}$ for DNAm sites located in islands and in sea obtained in peripheral blood lymphocytes (20) and outlined in the background section of this paper may have resulted from a) the inclusion of all DNAm sites rather than just those with a significant heritability estimate b) the use of the pedigree to estimate the contribution of the whole genome to phenotypic variance and/or c) bias due to un-modelled sources of environmental variation. In conjunction, CG content +/- 5KB of a TSS was inversely associated with the (genome-wide) heritability of gene expression measured in Peripheral Blood from 1,444 twin pairs (36).

The highest proportions of significantly heritable DNAm sites were located in intergenic regions as opposed to within the TSS of a gene or intragenic regions. Moreover, for DNAm sites within islands, being located in the TSS of a gene expressed in colon tissue led to a significantly lower probability of being heritable compared to being located within the TSS of a gene not expressed in colon tissue.

However, this pattern was not observed within the sea. This is suggestive of forces acting in a different manner at TSS within islands compared to at TSS within sea to maintain DNAm levels. Overall, the proportion of heritable DNAm sites was correlated with the average estimated genetic variance. A similar observation has been made for gene expression in Epstein-Barr virus transformed LCLs (37). In this case, the lower proportion of heritable DNAm sites observed for a contextual group(s) such as islands compared to other contextual group(s) such as the sea can in part be explained by a reduction in the measured genetic variance. Lower genetic variation could result in lower power to capture the true causative loci or it could be indicative of lower causal variation due to selective constraints.

DNAm level was slightly higher at the TSS of genes expressed in WCB and not LCM compared to those expressed in LCM. This result is consistent with the WCB samples being enriched for epithelial cells and a negative correlation between gene expression and DNA methylation level. However, the overall heritability of DNAm level at the TSS of genes expressed in WCB and not LCM compared to those expressed in LCM was similar. One possible explanation for these results is that in healthy colonic tissue the genes expressed in the colon epithelium are regulated by DNAm level in a similar fashion to the genes expressed in the WCB.

Finally, we have shown that genetic variants in genomic risk regions for CRC can affect DNAm level in healthy colon tissue and that overall DNAm sites within a risk region have similar overall heritability to DNAm sites outwith an identified risk region. In conjunction, we have replicated the previous finding that the CRC risk SNP rs4925386 effects DNAm level at cg15193198 and cg24112000.

We showed that when rs4925386 is excluded the regional genetic variation sufficiently captures the causal variation in DNAm level tagged by rs4925386. Moreover, rs4925386 alone does not capture all the genetic variance contributing to variation of cg24112000 and cg15193198 that is captured by the

regional heritability approach. This final result highlights the advantage of the regional heritability approach to capture the genetic effects on the phenotype, in this case DNAm level.

We have identified individual SNPs outwith +/- 1MB a DNAm site which affect DNAm level. However, studies of larger sample size are required to estimate the combined long-range effects (polygenic effect) of genetic variants on DNAm level with nominally unrelated individuals using the regional heritability method. This is because variance in the extent of identify by descent between nominally unrelated individuals is lower across the genome as a whole than it is within a small genomic region. However, in accordance with (22) we have shown that a small sample of nominally unrelated individuals can be used to estimate the genetic contribution of a genomic region to variation in DNAm level.

We have identified a subset of DNAm sites genome-wide and measured in healthy colon tissue that are influenced by the local genetic variation. Therefore, we have contributed to understanding healthy genetically influenced variation in DNAm level in colon tissue. A number of the DNAm sites which we report as heritable are located within CRC risk loci and thus have the potential to mediate genetic susceptibility to CRC. We expect further studies will focus on exploring a role for these DNAm sites in disease aetiology.

## Materials and Methods

### Samples

A total of 144 samples from normal colorectal tissue were obtained from Colombian patients diagnosed with either adenocarcinoma or adenomas of the colorectum. The study had ethical approval from the Ethics Board of The National Cancer Institute of Colombia, and participants gave informed written consent.

Similarly, 12 people undergoing colonoscopic examination at General University Hospital of Elche (Spain) but without adenocarcinoma or adenomas of the colorectum provided tissue samples from the colon. Written informed consent for inclusion in the study was obtained from every participating individual. The study was approved by the ethics committees of the General University Hospital of Elche.

### Phenotype QC

We assayed 144 samples for DNAm level at 485512 DNAm sites using the HM450K array. Within each colorectal tissue sample the two intensity values that correspond to the number of methylated and unmethylated copies of a DNAm were corrected for any variation that arose from non-specific binding. This background correction was applied by subtracting the median fluorescence measured by the control probes from the intensity values treating intensities measured in the two colour channels separately and using the Bioconductor package, 'lumi' (38). Subsequently, 9 samples for which the assaying process failed were identified. These samples had a low average intensity value (below 2500) measured in either or both of the colour channels and were removed. We examined the percentage of probes that were not detected above background levels of variation (P = 0.01) for each sample and found that no samples exceeded our threshold of 5% for exclusion. Two samples where the recorded sex of the individual did not match the sex estimated from the levels of DNAm measured on the X chromosome were removed. A total of 280,469 DNAm probes were removed because they contained a SNP within the target sequence or at the site of single base extension or they were deemed cross-reactive based on the work of (39) and the information provided

in the HM450K manifest file. Additionally, 354 probes were removed because they were not detected above background levels of variation for greater than 5% of samples (P<0.01). This resulted in 133 samples and 196080 autosomal DNAm probes left for downstream analysis. Colour bias was taken into account by comparing for all samples the within sample distribution of total intensities measured by the type I probes in the green channel to those measured in the red channel. A quantile normalization adjustment was applied within the Bioconductor package, 'lumi' (38), so that the intensity values measured in the two colour channels followed a similar distribution across and within individuals. We also applied a correction to account for technical variation due to the probe design type using the BMIQ algorithm (40).

We conducted analyses and reported levels of DNAm using the M-value scale. This scale reduces the dependence between the variance and mean of site-specific DNAm level that is observed on the beta scale (41). M-values are a logit transformation of the beta values and an M-value of 0 equates to a 50% level of methylation whereas a positive and a negative M-value relates to a greater and less than 50% methylation level respectively.

**Genotype QC**

We genotyped 468 samples at 958178 SNPs genome-wide using the Illumina HumanOmniExpress Exome Chip. We followed a standard quality control procedure (reviewed in (42)) using Plink (43). Four samples for which greater than 5% of SNPs did not genotype were excluded. Based on the application of four successive filters, SNPs were removed if 1) they failed to type in greater than 5% of samples or 2) if they were out of Hardy Weinberg Equilibrium (P<0.0001) or 3) if they had a MAF less than 0.01 or 4) if the rate of genotype failure was significantly different in cases and controls (P<0.00001). This procedure left a total of 682945 autosomal SNPs for analysis. Additionally, the inbreeding coefficient for each sample was calculated from SNPs along the X chromosome. This analysis revealed 3 samples recorded as female that were more inbred than expected (F > 0.98) and 10 samples recorded as male that were less inbred than expected (F <0.2). These samples were removed from subsequent analysis due to an assumed discrepancy between the recorded and observed identity.

This left 132 samples with quality genotype information that overlapped with the samples assayed for DNAm level and which passed the DNAm level quality control procedures.

**Identification of Genes Expressed in Colon Tissue**

Genes expressed in general colon tissue and specifically in colon epithelial cells were identified based on the analysis of normal tissue from biopsies of 12 people undergoing colonoscopic examination at General University Hospital of Elche (Spain). In order to separate epithelial specific expression, tissue samples were sliced with alternate slices assigned to the *whole tissue* and *epithelial* conditions. In the whole tissue condition combined slices for each individual were assayed for gene expression. In the epithelial condition we pooled epithelial cells, isolated using Laser Capture Microdisection (MMI CellCutPlus), from each slice for each individual and mRNA from the slices was amplified prior to being assayed for gene expression. The gene expression assay on the 24 samples was performed using the HumanHT-12 Expression BeadChip. Quality control indicated failure of four samples (one in the whole tissue and three in the epithelial condition) which were removed from subsequent analysis. We then identified for each condition mRNA probes which were detected above background ($P<0.01$) in more than 80% of samples, i.e., 9 or more of 11 samples and 8 or more of 9 samples for the whole tissue and epithelial conditions respectively. This yielded 9223 probes in the whole tissue and 4071 probes in the epithelial conditions. Probes were mapped to genes using the Illumina provided manifest file for the HumanHT-12 Expression BeadChip platform, yielding a list of 8114 genes expressed in general colon tissue and 3754 genes expressed in epithelium. As expected a majority of genes identified in the epithelial condition were also detected in the whole tissue which contains both the epithelial and other cells, with only 10 specific to the epithelial condition. This supports the view that the genes identified form subset of genes enriched for epithelial specific expression.

**Location of a DNAm site with Respect to a Gene**

A DNAm site was considered to be located within the TSS of a gene expressed in colon tissue if in the manifest file the site was recorded as being located within 200bp upstream of the TSS (TSS200) or, within 200-1500bp upstream of the TSS (TSS1500) of a gene in our list of expressed genes. All

other DNAm sites located within the TSS200 or TSS1500 region of a gene were considered as being located within a gene not expressed in colon tissue. Intragenic DNAm sites were those documented in the manifest file as located within the 5'UTR, 1st exon, gene body or 3'UTR. Intergenic DNAm sites were those not documented as residing within a gene. We applied successive filters in the aforementioned order so that each DNAm site fit into one of the four mutually exclusive categories.

**Statistical Analysis**

A WCB of healthy tissue was obtained from the ascending, transverse, descending, or sigmoid colon, cecum, rectum or region where the sigmoid colon joins the rectum from Colombian subjects who attended Colonoscopy examination and with diagnosis of hyperplasic polyp, adenoma, insitu carcinoma or carcinoma of the rectum or colon. In our analyses we included both diagnosis and biopsy location as explanatory variables. Biological differences between the right (proximal) and left (distal) colon have been identified and they include the tissue of developmental origin and manifestation of CRC (44). Therefore, we used the WCB location to define a new variable that indicated the location of the WCB with respect to left and right colon. The right colon included the cecum, ascending and transverse colon. The left colon included the descending and sigmoid colon, sigmoid-rectum union and the rectum. This variable was used as the explanatory variable to adjust for the effects of WCB location. We also accounted for sex, age and batch (plate) in our analyses following the work of others who have shown that these variables can affect DNAm level (Gibbs, Lam, Boks, Tapp). In addition, we fitted two genotype principal components to account for any stratification within our population sample. Results were practically identical when the analyses were done adjusting for sex and age.

The components of variance were estimated by fitting a mixed linear model using restricted maximum likelihood and the publically available software: REACTA (45). Consider $y$ a vector of measurements of DNAm level across all samples ($n$) for a single DNAm site, $\beta$ a vector measuring the effects sex and age and $X$ a design matrix mapping the appropriate explanatory variable to each sample. Then

with $W$ a matrix of standardized SNP genotypes from each sample and assuming a vector of SNP effects, $u \sim N(0, I\sigma_u^2)$ with $I$ a diagonal matrix and the random error $\varepsilon = N(0, I\sigma_\varepsilon^2)$ the model is defined as:

$$y = X\beta + Wu + \varepsilon$$

The heritability attributable to SNPs local to the DNAm site can be estimated from the following equation:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$$

Where $\sigma_g^2 = N\sigma_u^2$ and $N$ is the number of SNPs, and $A = WW'/N$ is a matrix of genetically derived relationships calculated from $N$ SNPs for *n* samples. The formula used for calculating the pairwise relationships from the SNP information can be found in VanRaden (46). Using a 1MB window either side of the DNAm we find that a total of between 1 and 3037 SNPs are included in the analysis (Figure 9). The null hypothesis that the heritability estimate was not significantly different from zero was tested with a Likelihood Ratio Test distributed as 50:50 mixture of Chi-squared distributions with 0 and 1 degrees of freedom (47). If $P < 0.05$ we rejected the null hypothesis and concluded that the DNAm site was heritable.

The SNP by SNP GWAS was conducted on residual DNAm level for each of 196080 DNAm sites using PLINK version 1.90 (43) and the --assoc command.

**Significance Testing of Proportions**

To test if two proportions are significantly different from one another we use the prop.test function in R (48). In brief, this function assumes that the two sample sizes are sufficiently large so that the distribution of the first proportion minus the second proportion is Gaussian. We apply a two-tailed test because we do not have prior expectation of the relative magnitudes of the two proportions being tested.

## Acknowledgements

This work was supported by National Cancer Institute of Colombia, Cancer Research UK [C12229/A13154] and The Roslin Institute Strategic Grant funding from the BBSRC. AT and CSH also acknowledges funding from the Medical Research Council Human Genetics Unit.

We thank ARK Genomics and the WTCRF Genetics Core at the University of Edinburgh for the DNA analyses. We thank James Prendergast for useful comments on the manuscript.

## Conflict of Interest Statement

The authors declare that there are no conflicts of interest

## References

1 Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315-322.

2 Cantone, I. and Fisher, A.G. (2013) Epigenetic programming and reprogramming during development. *Nat. Struct. Mol. Biol.*, **20**, 282-289.

3 Sasaki, H. and Matsui, Y. (2008) Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat. Rev. Genet.*, **9**, 129-140.

4 Weaver, J.R., Susiarjo, M. and Bartolomei, M.S. (2009) Imprinting and epigenetic changes in the early embryo. *Mamm. Genome.*, **20**, 532-543.

5 Bestor, T.H. (2000) The DNA methyltransferases of mammals. *Human Molecular Genetics*, **9**, 2395-2402.

6 Chen, T. and Li, E. (2006) Establishment and maintenance of DNA methylation patterns in mammals. *Curr. Top. Microbiol. Immunol.*, **301**, 179-201.

7 Esteve, P.O., Chin, H.G., Smallwood, A., Feehery, G.R., Gangisetty, O., Karpf, A.R., Carey, M.F. and Pradhan, S. (2006) Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. *Genes Dev.*, **20**, 3089-3103.

8 Wang, J., Hevi, S., Kurash, J.K., Lei, H., Gay, F., Bajko, J., Su, H., Sun, W., Chang, H., Xu, G. *et al.* (2009) The lysine demethylase LSD1 (KDM1) is required for maintenance of global DNA methylation. *Nat. Genet.*, **41**, 125-129.

9 Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., 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.

10 Lam, L.L., Emberly, E., Fraser, H.B., Neumann, S.M., Chen, E., Miller, G.E. and Kobor, M.S. (2012) Factors underlying variable DNA methylation in a human community cohort. *Proc. Natl. Acad. Sci. USA.*, **109**, 17253-17260.

11 Boks, M.P., Derks, E.M., Weisenberger, D.J., Strengman, E., Janson, E., Sommer, I.E., Kahn, R.S. and Ophoff, R.A. (2009) The Relationship of DNA Methylation with Age, Gender and Genotype in Twins and Healthy Controls. *PLoS ONE*, **4**, e6767.

12 Tapp, H.S., Commane, D.M., Bradburn, D.M., Arasaradnam, R., Mathers, J.C., Johnson, I.T. and Belshaw, N.J. (2013) Nutritional factors and gender influence age-related DNA methylation in the human rectal mucosa. *Aging Cell*, **12**, 148-155.

13 Lokk, K., Modhukur, V., Rajashekar, B., Martens, K., Magi, R., Kolde, R., Koltsina, M., Nilsson, T.K., Vilo, J., Salumets, A. *et al.* (2014) DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.*, **15**, r54.

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

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

16 Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, Shiao-Lin, Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J. *et al.* (2010) Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. *PLoS Genet.*, **6**, e1000952.

17 Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y. and Pritchard, J.K. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.*, **12**, 405-418.

18 Gervin, K., Hammerø, M., Akselsen, H.E., Moe, R., Nygård, H., Brandt, I., Gjessing, H.K., Harris, J.R., Undlien, D.E. and Lyle, R. (2011) Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res.*, 1813-1821.

19 Quon, G., Lippert, C., Heckerman, D. and Listgarten, J. (2013) Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic Acids Res*, **41**, 2095-2104.

20 McRae, A., Powell, J., Henders, A., Bowdler, L., Hemani, G., Shah, S., Painter, J., Martin, N., Visscher, P. and Montgomery, G. (2014) Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol.*, **15**, R73.

21 Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S. and Liu, C. (2010) Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. *The American Journal of Human Genetics*, **86**, 411-419.

22 Quon, G., Lippert, C., Heckerman, D. and Listgarten, J. (2013) Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic Acids Res.*, **41**, 2095-2104.

23 Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, **88**, 76-82.

24 Nagamine, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., Campbell, H., Wilson, J., Wild, S., Hicks, A.A. *et al.* (2012) Localising Loci underlying Complex Trait Variation Using Regional Genomic Relationship Mapping. *PLoS ONE*, **7**, e46501.

25 Uemoto, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Wilson, J.F., Rudan, I., Campbell, H., Hastie, N.D., Wright, A.F. *et al.* (2013) The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. *Front. Genet.*, **4**, 232.

26 Huels, D.J. and Sansom, O.J. (2015) Stem vs non-stem cell origin of colorectal cancer. *Br. J. Cancer*, **113**, 1-5.

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

28 Hindorff, L.A., MacArthur, J., Morales, J., A., J.H., Hall, P.N., Klemm, A.K. and Manolio, T.A., Available at: http://www.genome.gov/gwastudies.

29 Heyn, H., Sayols, S., Moutinho, C., Vidal, E., Sanchez-Mut, J.V., Stefansson, O.A., Nadal, E., Moran, S., Eyfjord, J.E., Gonzalez-Suarez, E. *et al.* (2014) Linkage of DNA methylation quantitative trait loci to human cancer risk. *Cell Rep.*, **7**, 331-338.

30 Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M. and Schubeler, D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457-466.

31 Cohen, N.M., Kenigsberg, E. and Tanay, A. (2011) Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell*, **145**, 773-786.

32 Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D/'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y. *et al.* (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, **466**, 253-257.

33 Saxonov, S., Berg, P. and Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. USA.*, **103**, 1412-1417.

34 Powell, J.E., Henders, A.K., McRae, A.F., Wright, M.J., Martin, N.G., Dermitzakis, E.T., Montgomery, G.W. and Visscher, P.M. (2012) Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome Res.*, **22**, 456-466.

35 Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A. *et al.* (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.*, **44**, 1084-1089.

36 Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H. *et al.* (2014) Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.*, **46**, 430-437.

37 Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217-1224.

38 Du, P., Kibbe, W.A. and Lin, S.M. (2008) lumi: a pipeline for processing Illumina microarray. *Bioinformatics.*, **24**, 1547-1548. doi: 1510.1093/bioinformatics/btn1224. Epub 2008 May 1548.

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

40 Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D. and Beck, S. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189-196.

41 Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L. and Lin, S.M. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

42 Weale, M. (2010) Quality control for genome-wide association studies. *Methods Mol. Biol.*, **628**, 341-372.

43 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**, 559-575.

44 Oh, S.W., Kim, Y.H., Choi, Y.S., Chang, D.K., Son, H.J., Rhee, P.L., Kim, J.J., Rhee, J.C., Yun, S.H., Lee, W.Y. *et al.* (2008) The comparison of the risk factors and clinical manifestations of proximal and distal colorectal cancer. *Dis. Colon Rectum*, **51**, 56-61.

45 Cebamanos, L., Gray, A., Stewart, I. and Tenesa, A. (2014) Regional heritability advanced complex trait analysis for GPU and traditional parallel architectures. *Bioinformatics*.

46 VanRaden, P.M. (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, **91**, 4414-4423.

47 Stram, D.O. and Lee, J.W. (1994) Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171-1177.

48 R Development Core Team. (2011). R Foundation for Statistical Computing, Vienna, Austria.

## Legends to Figures

**Figure 1: Distribution of Mean Site-specific DNAm level with respect to CpG density**

Methylation levels were measured on the M-value scale where a DNAm level of 0 can be interpreted as a 50% methylation level, a DNAm level < 0 and a DNAm level greater than > 0 indicate lower and greater than 50% methylation respectively. The majority of DNAm sites in islands exhibited a low average methylation level, which was in contrast to the majority of DNAm sites in low density CG regions (sea) being on average highly methylated.

**Figure 2: Mean Site-specific DNAm level as a function of distance from the edge of the island**

The 4000 BP region upstream (North) and downstream (South) of Islands was divided into bins of 100 BP. The average of the mean site-specific DNAm levels for DNAm sites residing within each bin is shown as a white circle enclosed by a line indicating +/- 2 standard error of the mean estimate. A shore is up to 2000 BP from an Island and a shelf is between 2000 and 4000BP from an Island.

**Figure 3: Distribution of mean site-specific DNAm level for eight contextual groups**

**Figure 4: Moments of the distributions of mean site-specific DNAm level for eight contextual groups**

**Figure 5: Distribution of the variance of each DNAm site used in analysis (n=196080)**

**Figure 6: Distribution of the Estimated heritability for DNAm sites significantly associated with local genetic variation.**

Each bar represents a range of 0.05.

**Figure 7: Number of SNPs +/- 1MB surrounding a DNAm site**

**Figure 8: Distribution of mean-site specific heritability for genes expressed in whole colorectal biopsies and colon epithelial cells**

DNAm sites have a significant heritability if P<0.05. Genes expressed in both the epithelial and whole colorectal biopsy (WCB) were removed from the WCB group for this analysis.

**Figure 9: The mean genetic variance and the proportion of heritable DNAm sites for the eight contextual groups.**

The x-axis value represents the proportion of heritable DNAm sites within each contextual group and the average genetic variance for each contextual group.

**Figure 10: Proportion of the regional heritability that can be explained by the top SNP association**

The variance explained by a SNP ($R^2$) divided by the regional heritability estimate for DNAm sites with a regional heritability estimate significant at P < 0.05. Only the most significant SNP within the local region was considered. In 4.2% of cases the SNP explained more variance in DNAm level than the region.

## Tables

**Table 1: Number of DNAm sites within six regions defined by physical distance from islands**

As described in the body of the manuscript, north and south shores encompass up to 2KB upstream and downstream of islands respectively. Regions 2-4KB upstream and downstream of islands were defined respectively as north and south shelves (27). Sea is any DNAm site not annotated as being located within an island, shelf or shore in the 450K manifest file.

| Genomic Context with Relation to CpG Density | Island | North Shore | South Shore | North Shelf | South Shelf | Sea | Total |
|---|---|---|---|---|---|---|---|
| Number of DNAm Sites | 74274 | 27405 | 21158 | 8323 | 7503 | 57417 | 196080 |

**Table 2: Number of DNAm sites within each of the eight contextual groups**

| Genomic Context | Island | Sea |
|---|---|---|
| TSS Expressed | 13838 | 2074 |
| TSS Not Expressed | 19052 | 7603 |
| Intragenic | 31356 | 30106 |
| Intergenic | 10028 | 17634 |
| Total | 74274 | 57417 |

**Table 3: Proportion of heritable DNAm sites and the corresponding mean heritability**

Overall there was a higher proportion of heritable DNAm sites located in the sea compared to islands. Additionally, there was a higher proportion of heritable DNAm sites located in intergenic regions compared to regions containing a TSS and intragenic regions. The average heritability estimates were similar across the contextual groups.

| | Island | | Sea | |
|---|---|---|---|---|
| | Proportion Heritable | Mean $h^2$ | Proportion Heritable | Mean $h^2$ |
| TSS Expressed | 0.066 | 0.275 | 0.117 | 0.288 |
| TSS Not Expressed | 0.084 | 0.281 | 0.117 | 0.293 |
| Intragenic | 0.089 | 0.283 | 0.100 | 0.290 |
| Intergenic | 0.129 | 0.308 | 0.123 | 0.303 |
| Total | 0.089 | 0.286 | 0.110 | 0.295 |

**Table 4: Significant Associations from SNP by SNP GWAS for 196080 DNAm sites**

The total number of significant associations (Count) and number of expected significant associations based on the specified threshold and the total number of SNPs tested (Count Expected) is reported. The number of DNAm sites with at least one significant association is given (Count DNAm Sites). SNPs were binned with respect to distance from the DNAm site (*cis* +/- 1MB and *trans* > +/- 1MB)

| | | Cis | | | Trans | |
|---|---|---|---|---|---|---|
| Threshold | Count | Count Expected | Count DNAm Sites | Count | Count Expected | Count DNAm Sites |
| $5*10^{-2}$ | 6,027,276 | 5,572,851 | 195,792 | | | |
| $5*10^{-4}$ | 184,978 | 55,729 | 55,532 | | | |
| $5*10^{-8}$ | 17,712 | 5.57 | 4,050 | 1,519,534 | 6,690 | 72544 |
| $5*10^{-12}$ | 9,495 | $5.57*10^{-4}$ | 1,412 | 236,577 | $6.69*10^{-1}$ | 16474 |
| $5*10^{-20}$ | 1,037 | $5.57*10^{-12}$ | 343 | 23,072 | $6.69*10^{-9}$ | 1786 |
| $5*10^{-40}$ | 18 | $5.57*10^{-32}$ | 8 | 706 | $6.69*10^{-29}$ | 47 |
| Max $R^2$ | 0.884 | | | 0.8717 | | |
| Min P-value | $1.40*10^{-62}$ | | | $1.96*10^{-58}$ | | |
| Total SNPs Tested | 111558037 | | | $1.34*10^{11}$ | | |

**Table 5: Effects of rs4925386 and local genetic variation on cg15193198 and cg24112000**

The regional heritability estimate ($\dot{h}^2_{r,g}$) for cg15193198 and cg24112000 including rs4925386. The estimate for the full model ($\dot{h}^2_{r,g}$ Full) and reduced model ($\dot{h}^2_{r,g}$ Reduced) were calculated from all SNPs +/-1MB of the DNAm site excluding rs4925386. The full model included fitting the genotypes at rs4925386 as a fixed effect. The effect of rs4925386 on DNAm level is reported as the addition of a single copy of the minor allele, Adenine.

| | $\dot{h}^2_{r,g}$ | $\dot{h}^2_{r,g}$ P-value | $\dot{h}^2_{r,g}$ Full | $\dot{h}^2_{r,g}$ Full P-value | $\dot{h}^2_{r,g}$ Reduced | $\dot{h}^2_{r,g}$ Reduced P-value | SNP Effect | SNP Effect SE |
|---|---|---|---|---|---|---|---|---|
| cg15193198 | 0.307 | $1.66 * 10^{-3}$ | 0.025 | 0.411 | 0.301 | $2.19*10^{-3}$ | -0.438 | 0.079 |
| cg24112000 | 0.625 | $5.51*10^{-12}$ | 0.281 | $1.44*10^{-2}$ | 0.629 | $8.47*10^{-12}$ | -0.576 | 0.090 |

## Abbreviations

DNAm, DNA methylation

LCM, Laser capture microdissection

WCB, whole colon biopsy

CRC, colorectal cancer

TSS, transcription start site

island TSS expressed, within an island and a TSS of a gene expressed in colon

island TSS not expressed, within an island and in a TSS of a gene not expressed in colon

island intragenic, within an island and intragenic

island intergenic, within an island and intergenic,

sea TSS expressed,within the sea and the TSS of a gene expressed in colon

sea TSS not expressed, within the sea and the TSS of a gene not expressed in colon

sea intragenic, within the sea and intragenic

sea intergenic, within the sea and intergenic