Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes

Citation for published version:
Nestor, CE, Ottaviano, R, Reddington, J, Sproul, D, Reinhardt, D, Dunican, D, Katz, E, Dixon, JM, Harrison, DJ & Meehan, RR 2012, 'Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes' Genome Research, vol. 22, no. 3, pp. 467-477. DOI: 10.1101/gr.126417.111

Digital Object Identifier (DOI):
10.1101/gr.126417.111

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Genome Research

Publisher Rights Statement:
Copyright © 2012 by Cold Spring Harbor Laboratory Press
Freely available online through the Genome Research Open Access option

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes

Colm E. Nestor,1,2 Raffaele Ottaviano,1,2 James Reddington,2 Duncan Sproul,1,2 Diana Reinhardt,2 Donncha Dunican,2 Elad Katz,1 J. Michael Dixon,1 David J. Harrison,1 and Richard R. Meehan1,2,3

1Breakthrough Breast Cancer Research Unit and Division of Pathology, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; 2MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh EH4 2XU, United Kingdom

The discovery of substantial amounts of 5-hydroxymethylcytosine (5hmC), formed by the oxidation of 5-methylcytosine (5mC), in various mouse tissues and human embryonic stem (ES) cells has necessitated a reevaluation of our knowledge of 5mC/5hmC patterns and functions in mammalian cells. Here, we investigate the tissue specificity of both the global levels and locus-specific distribution of 5hmC in several human tissues and cell lines. We find that global 5hmC content of normal human tissues is highly variable, does not correlate with global 5mC content, and decreases rapidly as cells from normal tissue adapt to cell culture. Using tiling microarrays to map 5hmC levels in DNA from normal human tissues, we find that 5hmC patterns are tissue specific; unsupervised hierarchical clustering based solely on 5hmC patterns groups independent biological samples by tissue type. Moreover, in agreement with previous studies, we find 5hmC associated primarily, but not exclusively, with the body of transcribed genes, and that within these genes 5hmC levels are positively correlated with transcription levels. However, using quantitative 5hmC-qPCR, we find that the absolute levels of 5hmC for any given gene are primarily determined by tissue type, gene expression having a secondary influence on 5hmC levels. That is, a gene transcribed at a similar level in several different tissues may have vastly different levels of 5hmC (>20-fold) dependent on tissue type. Our findings highlight tissue type as a major modifier of 5hmC levels in expressed genes and emphasize the importance of using quantitative analyses in the study of 5hmC levels.

[Supplemental material is available for this article.]

Typically, mammalian DNA methylation involves the covalent attachment of a methyl group to the 5-position of cytosine in a CpG dinucleotide by a member of the DNA methyltransferase (DNMT) family of enzymes. DNA methylation patterns are established in early development by the de novo methyltransferases, DNMT3A/B, and maintained after subsequent cell divisions by the maintenance methyltransferase, DNMT1 (Bird 2002; Wu and Zhang 2010). The ability to establish and maintain DNA methylation patterns is essential for normal development in mammals (Li et al. 1992). Specifically, DNA methylation is involved in the maintenance of genome integrity by silencing transposable elements (Walsh et al. 1998), the process of X-inactivation in females (Lock et al. 1987; Sado et al. 2000), regulation of allele-specific expression at imprinted loci (Plass and Soloway 2002), and may be involved in the regulation of single-copy gene expression (Borgel et al. 2010). The genomic distribution of CpG methylation in the human genome is highly non-random; whereas the majority of CpGs are methylated, regions of high CpG density, termed CpG islands, which often colocalize with gene promoters, are typically unmethylated. How this dichotomous pattern of CpG methylation is maintained is unclear.

In a groundbreaking study, the DNA of mouse Purkinje neurons and embryonic stem (ES) cells were found to contain significant levels of the modified base 5-hydroxymethylcytosine (5hmC) (Kriaucionis and Heintz 2009). A complementary study identified the ten-eleven translocation (TET) protein, TET1, as an iron- and α-ketoglutarate (α-KG)-dependent dioxygenase responsible for catalyzing the hydroxylation of 5mC to 5hmC (Tahiliani et al. 2009). Subsequent studies have reported significant levels of 5hmC in many other mouse and human tissues (Globisch et al. 2010; Szlagierszak et al. 2010; Jin et al. 2011) and confirmed TET2 and TET3 as possessing enzymatic properties similar to that of TET1 (Ito et al. 2010; Koh et al. 2011). Interestingly, the genes TET1 and TET2 were first identified as common targets of mutation in Acute Myeloid Leukaemia (AML). Recently, Figueroa et al. (2010) showed that mutation of the isocitrate dehydrogenase genes IDH1 and IDH2 led to the aberrant production of 2-hydroxoglutarate (2-HG), a metabolite that inhibits TET2 enzymatic activity, resulting in a hypermethylated promoter phenotype in AML tumors carrying IDH1/2 mutations. This hypermethylated promoter phenotype overlapped with AML tumors carrying a TET2 mutation only, highlighting a potential role for the TET family of proteins in the epigenetic dysregulation observed in many cancers (Figueroa et al. 2010). In addition, Tet1 knockdown in mouse ES cells resulted in de novo methylation of the transcription start site (TSS) of more than 100 genes, although the gain in 5mC was modest (Williams et al. 2011).

Many of the techniques commonly used to assay 5mC, including conventional methyl-sensitive restriction digest and the widely used technique of bisulfite sequencing, are incapable of distinguishing between 5mC and 5hmC (Huang et al. 2010; Jin et al. 2010; Nestor et al. 2010). As such, the discovery of 5hmC...
necessitates not only its characterization in mammalian cells, but also reevaluation of previously published 5mC data (Dahl et al. 2011). The recent development of 5hmC-specific antibodies and 5hmC-sensitive restriction digest-PCR (5hmC-qPCR) assays has allowed for determination of genome-wide patterns and locus-specific amounts of 5hmC levels, respectively (Davis and Vaisvila 2011; Jin et al. 2011).

The biological function of 5hmC is unknown. However, several very recent studies of the genome-wide distribution of both 5hmC and TET1-binding in mouse ES cells have begun to illuminate potential roles of both the mark and the proteins, which in some cases may not only be separate, but conflicting. A role for 5hmC as an intermediate in DNA demethylation has been widely postulated (Tahiliani et al. 2009; Wu and Zhang 2010). As 5hmC is poorly recognized by DNMT1 (Valinluck and Sowers 2007), methylation can be lost passively through successive cell divisions. However, the rapid loss of 5mC from the paternal pro-nucleus in the zygote and in primordial germ cells has long hinted at a process of active demethylation in mammals (Mayer et al. 2000; Oswald et al. 2000; Hajkova et al. 2002). It is hypothesized that conversion of 5mC to 5hmC may serve as the first step in such a pathway by providing a substrate for downstream repair pathways such as mismatch (MMR) or base-excision repair (BER), which ultimately replace 5hmC with cytosine (C) (Wu and Zhang 2010). In support of this hypothesis, two recent reports have shown that the rapid loss of 5mC from mouse paternal pro-nuclei is accompanied by an accumulation of genome-wide 5hmC (Iqbal et al. 2011; Wossidlo et al. 2011). However, the failure to find many of the predicted intermediates of an active oxidative demethylation pathway by mass-spectrometry analysis of normal mouse brain tissues challenges the existence of such a mechanism (Globisch et al. 2010). The observations that 5hmC is enriched in the gene body of active genes in mouse cerebellum (Song et al. 2010) and cannot be bound by transcriptionally repressive methyl-CpG binding domain (MBD) proteins in vitro (Valinluck et al. 2004; Jin et al. 2010) strongly suggest a role for 5hmC in the regulation of transcription. Gene-body 5hmC levels may regulate the transcription rate by modifying the accessibility of genic chromatin to transcriptional machinery or by inhibiting binding of repressive methyl-CpG binding proteins (MBDs). However, work on mouse ES cells suggests that the relationship between 5hmC and/or TET1 with transcription is far more complex (Pastor et al. 2011; Williams et al. 2011; Wu et al. 2011a,b). TET1 enrichment was often observed at the transcription start site (TSS) of genes with high CpG density (HCP) promoters marked by the bivalent histone signature of histone 3 lysine 27 trimethylation (H3K27me3) and histone 3 lysine 4 trimethylation (H3K4me3). This gene set is enriched for transcriptionally repressed developmental genes. Moreover, the repressive activity of TET1 may be independent of its catalytic role, because several genes up-regulated upon TET1 knockdown in WT ES cells were also up-regulated in ES cells lacking DNA methylation (Dnmt1/3a/b−/−) (Williams et al. 2011). Consistent with these observations, Wu and colleagues found 5hmC to be enriched in the bodies of highly expressed genes and at the TSS of transcriptionally inactive genes (Wu et al. 2011a). However, two simultaneous studies failed to report a significant association between genic 5hmC and transcription levels or RNA polymerase II activity (Pastor et al. 2011; Williams et al. 2011).

The majority of recent genome-wide studies of 5hmC and TET1 have used the same model system, mouse ES cells. Thus, the nature of locus-specific patterns of 5hmC within and between normal tissues remains unknown. Here, we use tiling microarrays and 5hmC-qPCR to investigate both global 5hmC content and locus-specific patterns of 5hmC in several normal human tissues. We find that unlike 5mC, global 5hmC content varies markedly between tissues and does not correlate with global 5mC content. Furthermore, the global 5hmC content of cell lines is markedly reduced compared with levels in the corresponding normal tissues. Significantly, we note that TET1/2/3 gene expression and consequently 5hmC content is rapidly and significantly reduced upon adaptation of cells from normal human tissue to cell culture. Using an 5hmC-specific antibody and tiling microarrays to profile locus-specific patterns of 5hmC, we found that 5hmC patterns are tissue-specific; allowing clustering of similar tissues based on 5hmC profiles alone. In addition, using RNA-seq data for normal human tissues, 5hmC was found to be enriched in the gene bodies of active genes, and the level of enrichment was positively correlated with gene expression levels. Significantly, however, we also find that for any given expressed gene, tissue type, not transcription level, is the major modifier of absolute levels of 5hmC.

Results

Global 5hmC content of human tissues is variable and does not correlate with total mC content or TET gene expression

PCR products of the mouse Ttx19.1 promoter in which each cytosine was either unmethylated (C), methylated (5mC), or hydroxymethylated (5hmC) were used to test the sensitivity and specificity of both the α-5mC and α-5hmC antibodies in immuno-dot-blot assays. Both antibodies exhibited specificity to their respective marks and no cross-reactivity was observed, even at higher concentrations of DNA (Fig. 1A). The dynamic range of detection of 5hmC implies that this assay can be usefully employed to determine differences in the 5hmC content of different genomic DNA samples.

Immuno dot-blots of DNA from eight normal human tissues and one human embryonic stem cell (hESC) line (Shf-6) revealed large inter-tissue variation in global 5hmC levels (Fig. 1B). In agreement with previous studies, human brain had the highest global levels of both 5hmC and 5mC (Song et al. 2010). Quantification of spot intensity by densitometric analysis further emphasized the scale of variation in global 5hmC between tissues. For example, levels in brain were about four times greater than those in the next highest tissue, breast, and more than 1000 times greater than the levels in blood or hESCs (Fig. 1C). The variation in inter-tissue global 5hmC levels contrasted strikingly with global 5mC levels, which varied little between tissues. Whereas the dot-blot assay could robustly detect the large inter-tissue differences observed in global 5mC levels, relative inter-tissue differences in 5mC were more difficult to quantify with confidence due to the small inter-tissue differences in global 5mC levels observed. To confirm the apparent lack of association between global mC and hmC levels determined by dot-blotting, the inter-tissue differences in 5hmC determined here were compared with previously published global 5mC levels quantified by high-performance liquid chromatography (HPLC) (Ehrlich et al. 1982; Weisenberger et al. 2005). No significant correlation was found between global 5hmC and 5mC levels (Spearman’s rho = −0.42; $P = 0.35$) (Fig. 1D). Interestingly, broadly similar patterns of inter-tissue global 5mC and 5hmC levels were conserved in mouse; brain and liver being highly enriched, whereas blood, spleen, and ES cells were relatively depleted of 5hmC (Fig. 1E).

To test the association between TET gene expression and global 5hmC content, we determined the relative expression level
of all three TET genes in 20 human tissues and human ES cells (Supplemental Fig. S2A). Each tissue cDNA sample was generated from RNAs pooled from at least three different individuals. The relative expression of TET2 and TET3 between tissues was highly similar to that reported in mouse (Ito et al. 2010). Human ES cells are distinguished by their relative lack of TET2 and TET3 expression, suggesting that TET1 may be the primary dioxygenase responsible for generation of 5hmC in these cells. However, human TET1 expression was just twofold higher in ES cells compared with average expression in somatic tissues. This is markedly lower than the 10-fold difference in Tet1 expression observed between ES cells and somatic tissues in mouse (Ito et al. 2010). Our observation of TET1 expression in human tissues is consistent with a previous analysis of TET1/2/3 gene expression in 24 human tissues that reported TET1 expression in half (12/24) of the tissues assayed (Lorsbach et al. 2003). Interestingly, TET1/2/3 gene expression did not correlate with relative global 5mC content (Supplemental Fig. S2B).

Cell culture results in a dramatic reduction of global 5hmC levels

Epigenetic dysregulation, including promoter hypermethylation and genome-wide hypomethylation, is a hallmark of many cancers (Esteller 2007). Moreover, several recent studies have reported a direct link between TET inhibition and aberrant promoter hypermethylation in AML (Figueroa et al. 2010; Xu et al. 2011). Thus, we determined if global 5hmC levels in cell culture models of several human cancers differed from that observed in their corresponding non-cancerous tissues of origin.

We observed much lower global 5hmC in eight breast cancer cell lines and one primary human mammary epithelial cell (HMEC) line relative to normal breast tissue (Fig. 2A,B). This reduction of global 5hmC in human breast cancer cell lines was mirrored by significantly reduced levels of transcripts for all three TET genes relative to levels in normal breast tissue (Fig. 2C). A similar reduction in global 5hmC levels was observed in both colon and liver cancer cell lines relative to levels in the corresponding non-cancerous tissues (Fig. 2D). Interestingly, DNA obtained from a primary HMEC line 1 wk after establishment from normal breast tissue also showed a dramatic reduction in both global 5hmC content and TET gene expression, suggesting that conversion to cell culture may be sufficient to effect a global reduction in 5hmC levels (Fig. 2A–C). This marked reduction in TET1/2/3 gene expression in culture was not observed between normal breast tissue and breast tumors in three previously published independent microarray studies, suggesting that the observed TET gene downregulation in culture may not be a cancer-associated process.
To further investigate this hypothesis, we cultured cells from normal human breast tissue and analyzed the changes in both global 5mC and 5hmC levels over time. Rapid loss of 5hmC was apparent after 1 wk in culture (1 passage; 7 d), and global 5hmC levels continued to decrease with each passage of the cells. In stark contrast, global 5mC levels remained unchanged, suggesting that the loss of 5hmC was not solely due to a concomitant loss of 5mC (Fig. 2E).

Locus-specific mapping of 5hmC in DNA from normal human tissues

The marked difference in global 5hmC levels observed between normal human breast tissue and analyzed the changes in both global 5mC and 5hmC levels over time. Rapid loss of 5hmC was apparent after 1 wk in culture (1 passage; 7 d), and global 5hmC levels continued to decrease with each passage of the cells. In stark contrast, global 5mC levels remained unchanged, suggesting that the loss of 5hmC was not solely due to a concomitant loss of 5mC (Fig. 2E).

We developed a combination of tiling microarrays and 5hmC-sensitive restriction digest-qPCR (hmC-qPCR) to determine locus-specific 5hmC patterns in several normal human tissues. As recent studies have revealed that both intra- and intergenic regions of the mouse genome can contain high levels of 5hmC, we used tiling microarrays covering the ENCODE regions selected for their biological significance, including the HOXA cluster, HBB, CFTFR, and APO cluster. The features represented on each array are listed in Supplemental Table S3. Using a candidate approach, we identified the promoter of TEX14 and a differentially methylated region (DMR) of the H19 gene as enriched for 5hmC, whereas LINE-1 elements and the promoter of GAPDH were relatively depleted. qPCR of these loci were used to validate all hydroxymethylated DNA immuno- precipitation (5hmC-IP) assays (Supplemental Fig. S4A).

Unsupervised hierarchical clustering of 5hmC enrichment values [log2(input/IP)] for all 72,000 probes on each array clustered samples by tissue type, confirming that patterns of 5hmC are tissue-specific and that inter-tissue variation in 5hmC patterns is greater than inter-individual variation (Fig. 3A,B).

Peaks of 5hmC enrichment were defined as any five consecutive probes in which a minimum of four probes had an enrichment score above the 75th percentile, yielding a false discovery rate (FDR) <0.05. We term probes assigned to 5hmC peaks "peak-probes." Subsequently, using gene models from the RefSeq database, each probe was classified according to its genomic location—
promoter, intragenic, downstream, or intergenic, as illustrated in Figure 4A (upper panel). Note that because repetitive sequences are typically not represented on tiling microarrays, only the non-repetitive portion of each genic location is assessed here. Consistent with previous studies in mouse ES cells, the genic distribution of peak probes differed significantly from that expected by chance ($\chi^2$-test; $P < 0.001$) being enriched in intragenic regions and promoters (Fig. 4A, lower panel). However, it is interesting to note that $\sim 25\%$ of ShmC peaks are not located within annotated genes and their proximal flanking sequences. To further investigate the observed association of ShmC enrichment and gene bodies, we analyzed the overlap between ShmC peak-probes and previously published total RNA-sequencing reads for normal human brain, breast, colon, liver, and testis (Ramskold et al. 2009). We find that peak-probes preferentially colocalize with RNA-seq reads in all tissues (Fig. 4B) and that the ShmC level of probes is positively correlated with transcription, as measured by the number of overlapping RNA-seq reads (Fig. 4C). Interestingly, although the association of ShmC levels with gene transcription within tissues is striking at some loci (Fig. 4D; Supplemental Fig. S5), it does not explain the patterns of ShmC observed at many other loci (Supplemental Fig. S6). Indeed, although significantly correlated, transcription (number of overlapping reads) explained <1% (i.e., brain; $r^2 = 0.0025$, $P = 0$) in the variation in ShmC values in each tissues assayed.

As tiling microarrays are limited in their ability to represent non-unique sequences, we determined the ShmC enrichment of several repetitive elements by ShmC-IP followed by qPCR. SINE elements (Alu), LINE-1 elements, and satellite sequences (Sat2 and Sat6) were all relatively depleted of ShmC in both normal tissues and cancer cell lines (Supplemental Fig. S7A). Thus, whereas much of the total 5mC content of a normal human genome is found in the repetitive fraction, ShmC appears largely occluded from repeat sequences. As “repeats” may comprise $>50\%$ of the human genome and are usually methylated in normal somatic tissue, the relative lack of global ShmC compared with global 5mC in reports for most normal mouse tissues may simply reflect a restriction of ShmC to unique, transcribed sequences (Weisenberger et al. 2005; Kriaucionis and Heintz 2009; Globisch et al. 2010).

**Figure 4.** Genomic patterns of 5hmC enrichment are tissue-specific. (A) Shown are the patterns of ShmC enrichment [log$_2$(input/IP)] across the HOXA cluster for multiple replicates of multiple tissues. (B) A dendrogram derived from unsupervised hierarchical consensus clustering of 14 human DNA samples based on ShmC enrichment levels for all 72,000 probes on each tiling microarray. Samples cluster by tissue type. AU (approximately unbiased) $P$-value of robustness of each cluster; (**$P > 0.01$; (***)$P > 0.001$.

**Table 1.** Differences in the percentage of 5hmC probes within annotated genes in normal mouse tissues.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Genic</th>
<th>Non-genic</th>
<th>Total</th>
<th>Genic/Non-genic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td>67%</td>
<td>45%</td>
<td>74%</td>
<td>1.48</td>
</tr>
<tr>
<td>Liver</td>
<td>64%</td>
<td>44%</td>
<td>70%</td>
<td>1.44</td>
</tr>
<tr>
<td>Colon</td>
<td>65%</td>
<td>45%</td>
<td>70%</td>
<td>1.45</td>
</tr>
<tr>
<td>Breast</td>
<td>63%</td>
<td>45%</td>
<td>68%</td>
<td>1.39</td>
</tr>
<tr>
<td>Testis</td>
<td>64%</td>
<td>45%</td>
<td>69%</td>
<td>1.42</td>
</tr>
</tbody>
</table>

**Figure 3.** Genomic patterns of 5hmC enrichment are tissue-specific. (A) Shown are the patterns of 5hmC enrichment [log$_2$(input/IP)] across the HOXA cluster for multiple replicates of multiple tissues. (B) A dendrogram derived from unsupervised hierarchical consensus clustering of 14 human DNA samples based on 5hmC enrichment levels for all 72,000 probes on each tiling microarray. Samples cluster by tissue type. AU (approximately unbiased) $P$-value of robustness of each cluster; (**$P > 0.01$; (***)$P > 0.001$.

**Tissue type, not transcription level, is the major modifier of ShmC levels in genes**

Whereas antibody-based enrichment of ShmC-containing DNA fragments allows for characterization of genome-wide ShmC patterns, it is difficult to relate the enrichment values obtained to absolute levels of ShmC, which can also be compromised by the effects of DNA sequence composition on antibody affinity. We used ShmC-sensitive restriction digest-qPCR (hmC-qPCR) to determine absolute ShmC levels at 12 loci across (1) the HOXA cluster, which shows marked differences in ShmC profiles between tissues, and (2) the H19 locus, which is among the most enriched regions assayed in all tissues (Supplemental Figs. S5, S6). An outline of the assay and sample calculations are presented in Supplemental Figure S1. We tested the accuracy of the assay by spiking each genomic sample with synthetic 100-bp DNA templates containing a single MspI site in which the internal cytosine was either unmodified (C), methylated (5mC), or hydroxymethylated (ShmC). Different ratios of each template type were added to different reactions. The expected %ShmC and experimentally determined %ShmC levels were highly significantly correlated (Pearson’s correlation; $r = 0.96$, $P > 0.0001$) over 10 independent experiments (Supplemental Fig. S8A). In addition, technical and biological replicates of both liver and brain samples were highly significantly correlated over seven loci tested, showing that the results obtained are highly reproducible (Supplemental Fig. S8B,C).

Marked variation in locus-specific ShmC content was observed between tissues; whereas the ShmC content of loci in the gene body of H19 varied between 37% and 70% in brain DNA, values of just 0.9%–4.3% were observed in blood (Table 1). In agreement with values obtained from tiling microarrays, each tissue had consistently higher levels of ShmC at loci in the H19/IGF2 region than those in the HOXA cluster, and genic loci in both regions showed higher levels of ShmC than non-genic loci (Supplemental Fig. S9A,B). However, no significant correlation between ShmC enrichment values determined on the microarray and those determined by ShmC-qPCR was observed over the 12 loci tested. The absence of significant correlation is not entirely unexpected given that ShmC enrichment values on the microarrays were determined by immuno-precipitation of large DNA fragments (200–
1000 bp) potentially containing numerous CpG dinucleotides, whereas 5hmC-qPCR evaluates a single MspI locus within that fragment, the 5hmC content of which may differ markedly from that of the fragment as a whole. As only 12 loci were tested and considerable variation between the microarray and 5hmC-qPCR was expected, we hypothesized that assaying additional loci would result in significant association between the two techniques. Ten additional loci were assayed by 5hmC-qPCR in the brain and liver samples. The values from the microarray and 5hmC-qPCR were now significantly correlated for liver (Pearson’s correlation coefficient; $r = 0.53$, $P = 0.01$) and approaching significance for brain (Pearson’s correlation coefficient; $r = 0.42$, $P = 0.051$) samples (Supplemental Fig. S8D).

Unsupervised hierarchical clustering of 5hmC levels at 12 loci in six tissues resulted in grouping the samples into two clusters corresponding to loci with high levels of 5hmC (the “HIGH” cluster; median 5hmC content = 17.2%) and those with lower levels of 5hmC (the “LOW” cluster; median 5hmC content = 2.2%). Consistent with the results obtained from the microarrays, the HIGH cluster consisted entirely of genic MspI sites, primarily from the H19/IGF2 locus, whereas the LOW cluster consisted of MspI sites from the HOXA cluster and non-genic MspI sites from...
Table 1. ShmC content (%) at 12 loci in six normal human tissues determined by ShmC-qPCR (EpiMark, NEB)

<table>
<thead>
<tr>
<th>Locus</th>
<th>Brain A</th>
<th>Breast A</th>
<th>Liver A</th>
<th>Placenta A</th>
<th>Testis B</th>
<th>Blood A</th>
</tr>
</thead>
<tbody>
<tr>
<td>H19 genic 3</td>
<td>69.7</td>
<td>22.3</td>
<td>24.7</td>
<td>2.8</td>
<td>14.1</td>
<td>1.3</td>
</tr>
<tr>
<td>H19 genic 4</td>
<td>47.8</td>
<td>37.4</td>
<td>10.6</td>
<td>18.7</td>
<td>32.0</td>
<td>4.3</td>
</tr>
<tr>
<td>IGF2 genic 1</td>
<td>43.5</td>
<td>14.4</td>
<td>16.8</td>
<td>27.0</td>
<td>47.1</td>
<td>2.7</td>
</tr>
<tr>
<td>H19 genic 1</td>
<td>36.7</td>
<td>11.8</td>
<td>14.8</td>
<td>13.0</td>
<td>10.6</td>
<td>0.9</td>
</tr>
<tr>
<td>HOXA6 intron</td>
<td>36.2</td>
<td>59.2</td>
<td>27.2</td>
<td>17.5</td>
<td>15.4</td>
<td>1.7</td>
</tr>
<tr>
<td>HOXA7 intergenic</td>
<td>10.6</td>
<td>3.7</td>
<td>10.2</td>
<td>7.4</td>
<td>5.6</td>
<td>0.0</td>
</tr>
<tr>
<td>HOXA6 exon</td>
<td>9.1</td>
<td>3.9</td>
<td>4.3</td>
<td>7.4</td>
<td>1.4</td>
<td>-0.4</td>
</tr>
<tr>
<td>H19 promoter</td>
<td>6.8</td>
<td>16.5</td>
<td>12.2</td>
<td>0.3</td>
<td>4.5</td>
<td>1.4</td>
</tr>
<tr>
<td>HOXA7 intron</td>
<td>2.6</td>
<td>8.8</td>
<td>3.7</td>
<td>10.1</td>
<td>4.5</td>
<td>-</td>
</tr>
<tr>
<td>HOXA7 intergenic</td>
<td>0.8</td>
<td>4.2</td>
<td>1.1</td>
<td>2.9</td>
<td>0.5</td>
<td>2.2</td>
</tr>
<tr>
<td>HOXA7 exon</td>
<td>-2.8</td>
<td>2.9</td>
<td>0.1</td>
<td>0.4</td>
<td>1.9</td>
<td>0.3</td>
</tr>
</tbody>
</table>

All values are single measurements determined by ShmC-qPCR (EpiMark, NEB). See Supplemental Figure S1 for experimental outline.

*Negative values were artificially set to 0.1 in analyses and plotting.

The discovery of 5-hydroxymethylation of cytosine as an abundant epigenetic mark in mammalian genomes has begun to redefine the field of DNA methylation and our interpretation of DNA methylation data past and present, most of which does not distinguish between canonical 5mC and ShmC (Nestor et al. 2010). Using immuno-precipitation of DNA with a ShmC-specific antibody, a plethora of recent studies have begun to describe the genomic location of ShmC and the potential function of both the mark and the TET enzymes that catalyze its generation from 5mC (Ficz et al. 2011; Pastor et al. 2011; Williams et al. 2011; Wu et al. 2011a,b). All have reported the preferential localization of both TET1 and ShmC in mouse ES cells; however, any correlation between ShmC and transcription levels is less clear. Unexpectedly, a small set of genes were up-regulated in knockdowns of TET1 in both wild-type mouse ES cells and mES cells lacking DNA methylation, suggesting a role for TET1 in transcriptional repression, which is independent of its catalytic activity. However, most recent genome-wide studies have performed very similar experiments, and all in the same model system, namely, mouse ES cells. A single study of human brain frontal lobe tissue similarly reported an association of ShmC with genic regions and a positive correlation between gene-body ShmC enrichment and transcription level (Jin et al. 2011). However, because this study examined only one tissue type and was restricted to promoter regions of genes, our knowledge of ShmC patterns within and between normal mammalian tissues remains basic.

Here, we present the first systematic study of both global and locus-specific patterns of ShmC in normal human tissues and both cancer- and non-cancer-derived cell lines. Importantly, in addition to immuno-precipitation-based assays of ShmC content, we use sensitive ShmC-qPCR to quantify absolute levels of ShmC at several loci in numerous tissues. We report large differences in global ShmC in normal human tissues, consistent with reports of global ShmC in mouse tissues (Kriaucionis and Heintz 2009). The inter-tissue difference in global ShmC is striking, brain DNA containing times the ShmC of blood DNA. Indeed, although we show that transcription level correlates weakly with ShmC levels within tissues, tissue type, and thus global ShmC content is a far more powerful predictor of the ShmC content in genes (Fig. 5C). Our failure to find significant correlation between gene expression and ShmC content between tissues is surprising (Fig. 5D) and suggests that factors other than transcription are the major modifiers of genic ShmC content. In addition, because the observed reduction in ShmC across all tissues was broadly conserved at all loci tested, inter-tissue differences in global ShmC are most likely due to a largely equal reduction in ShmC content across the genome, as opposed to disproportionate loss at certain genomic elements (i.e., repeats).

ShmC is widely presumed to be an intermediate in a process of DNA demethylation, either active or passive. Assuming a model of passive demethylation, ShmC loss resulting from failure to maintain the mark after DNA replication, one would predict that tissues with a relatively low proliferation rate (i.e., brain) would...
have the highest levels of 5hmC and highly proliferative tissues and cell lines would contain the least, as is broadly observed (Fig. 1C). Indeed, adaptation of normal breast cells to culture conditions did result in a decrease in 5hmC levels over time, consistent with an increase in passive loss of 5hmC due to an increased proliferation rate in culture (Fig. 2E). However, a simultaneous down-regulation of TET1/2/3 gene expression was also observed, suggesting that the reduction in 5hmC was due to both passive loss and reduced production of 5hmC (Fig. 2C,E). Irrespective of the mechanism of TET repression and 5hmC loss in culture, this finding highlights the marked difference in 5hmC biology between cell lines and their tissues of origin. Further study of the loss of 5hmC upon transformation of tissues to cell culture may offer a useful tool in dissecting 5hmC biology in mammals.

We also report a dramatic reduction of global 5hmC levels and TET1/2/3 gene expression in breast cancer cell lines compared with normal breast tissue (Fig. 2A–D). This observation is particularly interesting, given that breast cancer cell lines possess extensive aberrant promoter hypermethylation, and these promoters are enriched for CpG Island promoters marked simultaneously by H3K4me3 and H3K27me3 (bivalent promoters) in ES cells, the same class of genes identified as being marked by TET1 occupancy in mouse ES cells (Ohm et al. 2007; Sproul et al. 2011; Williams et al. 2011).

In summary, our results reveal that tissue type is a major modifier of both global and locus-specific 5hmC at genes in normal human tissues, suggesting that the functional importance of 5hmC varies between tissues. We also show that cell culture is refractory to both TET1/2/3 gene expression and maintenance of global 5hmC content. Our study emphasizes the importance of the model system (tissue or cell line) used to study 5hmC and the pressing need for more practicable quantitative assays of 5hmC.

Figure 5. Tissue type is a major modifier of 5hmC levels in genes. (A) Unsupervised hierarchical clustering of 12 loci by absolute 5hmC levels results in two groups containing high (HIGH) and low (LOW) levels of 5hmC, respectively, for each indicated tissue. (Red boxes) H19/IGF2 loci; (blue boxes) HOXA loci. (B) Box-plot of 5hmC (%) levels of all five loci in the HIGH cluster for each tissue shows that the range of 5hmC levels at HIGH loci varies markedly between tissues. (C) The relative global 5hmC content (Fig. 1C) correlates with average 5hmC (%) content of each tissue as determined by 5hmC-qPCR. (D) Inter-tissue differences in 5hmC levels do not correlate with inter-tissue expression levels. Average 5hmC (%) content of genic loci of H19 (upper panel) and IGF2 (lower panel) in each indicated tissue is plotted against expression of probes (H19 = 1 probe; IGF2 = 2 probes) for each gene obtained from the GNF expression database. The GNF expression database does not contain data for breast.
Methods

Ethics statement

The use of human breast materials from the Edinburgh Breast Unit at the Western General Hospital was approved by the Lothian Research Ethics Committee (08/S1101/41).

DNA, RNA, and tissue samples

Normal human tissue DNA samples were purchased from AMS Biotechnology Ltd. unless stated otherwise. Normal human tissue RNA samples were purchased from Applied Biosystems. SHEF-6 human embryonic stem cell DNA and RNA was a kind gift from Dr. David Hay (Center for Regenerative Medicine, Edinburgh, UK) (Aflatoonian et al. 2010). Clinical specimens of normal human breast tissue were obtained through the Edinburgh Experimental Cancer Medicine Center. Sample details are given in Supplemental Table S1. DNA and RNA were isolated from cell lines using the Qiagen AllPrep DNA/RNA/Protein Mini Kit.

Cell culture

All human breast cancer cell lines were cultured as per Neve et al. (2006). HepG2, HCT116, and SW480 cells were cultured in DMEM, McCoy’s 5A and Leibovitz’s L-15 medium, respectively, supplemented with 10% fetal calf serum (FCS). Human mammary epithelial cell (HEMRC) lines were established from normal human breast tissue as previously described (Holliday et al. 2009). Subsequently, HMECs were maintained in CnT22 medium (CellnTEC) supplemented with 10% FCS.

DNA dot-blotting

DNA samples were added to denaturation buffer (0.4 mM NaOH, 10 mM EDTA) and denatured for 10 min at 100°C. Samples were rapidly chilled for 5 min on wet ice and then applied to a positively charged nylon membrane under vacuum using a 96-well Dot Blot Hybridisation Manifold (Harvard Apparatus Limited). The membrane was washed twice in 2× SSC buffer, UV-cross-linked, and dried for 1 h at 70°C. Duplicate membranes were probed with antibodies specific to 5mC (Eurogentech; dilution factor 1:2000) and 5hmC (Active Motif; dilution factor 1:8000). To control for loading, duplicate membranes were probed with a rabbit polyclonal antibody raised against single-stranded DNA (Demeditec Diagnostics). Subsequently, membranes were probed with either a rabbit (α-5mC and α-ssDNA membranes) or mouse (α-5hmC membranes) IgG antibody conjugated to horseradish peroxidase (HRP). Following treatment with enhanced chemiluminescence (ECL) substrate, membranes were scanned on an ImageQuant LAS 4000 (GE Healthcare) imaging station. Spot intensity was quantified using ImageJ image processing and analysis software (NIH).

Immunoprecipitation and genomic mapping

5mC- and 5hmC-containing DNA fragments

Genomic DNA (5mC/IP: 2.5 μg in 450 μL of TE), sonicated to yield a fragment distribution of ~300–1000 bp, was denatured by incubation for 10 min at 100°C. Samples were rapidly chilled on wet ice. At this point, 45 μL (10%) of denatured sample was removed and saved as input, and 45 μL of 10× IP buffer (100 mM Na2HPO4 at pH 7.0 [mono and dibasic], 1.4 M NaCl, 0.5% Triton X-100) and 1 μg of α-5hmC (ActiveMotif; #39769) antibody were added to the remaining sample. Samples were incubated overnight at 4°C with gentle agitation. Then, 40 μL of magnetic beads (Dynabeads Protein G; Invitrogen) in 1× IP buffer was added to each sample to allow magnetic separation of the antibody from the unbound DNA using a magnetic tube rack. Samples were incubated for 1 h at 4°C with gentle agitation. Beads were collected with a magnetic rack and washed with 1000 μL of 1× IP buffer for 10 min at room temperature with gentle agitation; washing was repeated three times. Beads were collected with a magnetic rack and resuspended in 250 μL of digestion buffer (50 mM Tris at pH 8.0, 10 mM EDTA, 0.5% SDS) followed by addition of 10 μL of proteinase K (20 mg/mL; Roche Applied Science) and incubation for 1.5 h at 52°C with constant shaking (~800 rpm). Finally, beads were removed using a magnetic rack, and DNA was purified from the remaining sample using a QIAquick PCR Purification Kit (Qiagen), eluting in a final volume of 40 μL of dH2O. Inputs were also purified using a QIAquick PCR Purification Kit and eluted in 40 μL of dH2O.

Subsequently, 10 ng of input and IP DNA was subjected to whole genome amplification (WGA) using the GenomePlex Complete Whole Genome Amplification Kit (Sigma-Aldrich) as per the manufacturer’s instructions. Amplified DNA was run on a 1.2% agarose gel to confirm consistency of fragment size between samples. Subsequently, amplified DNA samples were Cy5- (IP) or Cy3- (Input) labeled by random priming using the Dual-Color DNA Labeling Kit (NimbleGen). Labeled samples were applied to a Human DNA Methylation 4x72K ENCODE HG17 Targeted Tiling Array, which tiles the biologically significant ENCODE regions (EN001–EN014), and hybridized overnight at 42°C. Slides were washed and scanned as per the NimbleGen protocol.

Analysis of microarray data

All analysis of NimbleGen microarray data was performed using custom-written scripts implemented in the statistical programming language R. Due to differences in the saturation kinetics of the dyes used to label the input and IP samples, the absolute enrichment values of a probe (M) often show a dependency on the average intensity of the probe across the input and IP (A). To correct for this bias, each array was subjected to locally weighted scatterplot smoothing (Loess normalization) (Cleveland 1979). Subsequently, scale normalization was used to normalize values of M between arrays of replicate samples. Finally, the data were smoothed using a running median in a sliding window of 13 probes. A peak of 5hmC was defined as any region of five consecutive probes in which at least four probes had an enrichment value greater than the 75th percentile. This definition of a peak resulted in a false discovery rate of 0.017. All R-scripts are available upon request. Although the microarrays were designed against the hg17 version of the human genome assembly, all coordinates were mapped to version hg18 to allow comparison with RNA-sequencing data. Conversion of genomic coordinates between genome assemblies was performed using the “liftOver” tool downloaded from the UCSC Genome Browser.

Previously published RNA-sequencing data for normal human breast, brain, colon, liver, and testis were downloaded from the UCSC Genome Browser (Ramskold et al. 2009). RNA-seq data were used in their normalized form expressed as reads per million (RPKM). Probes were designated as in a transcribed region if they overlapped a sequencing read by one or more base pairs. Where the density of reads overlapping a probe varied across the length of the probe, the average density was used.

Absolute quantification of 5-hmC and 5-mC levels at specific CpG dinucleotides

Absolute levels of 5hmC, 5mC, and C at specific MspI sites were quantified using the EpiMark 5-hmC and 5-mC Analysis Kit.

Tissue specificity of 5hmC content and patterns
RNA extraction, cDNA synthesis, and quantitative reverse transcriptase PCR

RNA was extracted from cell lines using TRIzol reagent (Invitrogen) or an RNaseasy Mini kit (QiAGEN) according to the manufacturers' instructions. A Superscript II Reverse transcriptase kit (Invitrogen) was used to make complementary DNA from 500 ng of total RNA. All qRT-PCRs were carried out at an annealing temperature 58°C in a LightCycler 480 Real-Time PCR System (Roche Applied Science). Primer sequences are given in Supplemental Table S2B.

Data access

All microarray data have been submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE33219.

Acknowledgments

We thank Tobias Straub and Johannes Soding for R-scripts and advice concerning the analysis of NimbleGen tiling array data. Thanks to Dr. James Hackett for help with development of the 5mC-IP technique and advice and Dr. John Thomson for helpful discussions. Thanks to Lorna Renshaw (Edinburgh Breast Unit) for assistance with obtaining breast materials. We acknowledge the support of Cancer Research UK for the Edinburgh Cancer Research Centre. This work was funded by the Breakthrough Breast Cancer and the Medical Research Council.

References


Nestor et al. 476 Genome Research www.genome.org


Received May 18, 2011; accepted in revised form November 10, 2011.