

# The Dynamics and Prognostic Potential of DNA Methylation Changes at Stem Cell Gene Loci in Women's Cancer

Joanna Zhuang<sup>1,2</sup>, Allison Jones<sup>1</sup>, Shih-Han Lee<sup>1</sup>, Esther Ng<sup>2</sup>, Heidi Fiegl<sup>3</sup>, Michal Zikan<sup>4</sup>, David Cibula<sup>4</sup>, Alexandra Sargent<sup>5</sup>, Helga B. Salvesen<sup>6,7</sup>, Ian J. Jacobs<sup>1,5</sup>, Henry C. Kitchener<sup>5</sup>, Andrew E. Teschendorff<sup>2\*</sup>, Martin Widschwendter<sup>1\*</sup>

**1** Department of Women's Cancer, University College London Elizabeth Garrett Anderson Institute for Women's Health, London, United Kingdom, **2** Statistical Genomics Group, University College London Cancer Institute, London, United Kingdom, **3** Department of Gynaecology and Obstetrics, Innsbruck Medical University, Innsbruck, Austria, **4** Oncogynecologic Center, Department of Obstetrics and Gynaecology, Charles University Prague—First Faculty of Medicine and General Faculty Hospital, Prague, Czech Republic, **5** School of Cancer and Imaging Science, Manchester Academic Health Science Centre, Central Manchester University Hospitals NHS Foundation Trust, Manchester, United Kingdom, **6** Department of Obstetrics and Gynaecology, Haukeland University Hospital, Bergen, Norway, **7** Department of Clinical Medicine, University of Bergen, Bergen, Norway

## Abstract

Aberrant DNA methylation is an important cancer hallmark, yet the dynamics of DNA methylation changes in human carcinogenesis remain largely unexplored. Moreover, the role of DNA methylation for prediction of clinical outcome is still uncertain and confined to specific cancers. Here we perform the most comprehensive study of DNA methylation changes throughout human carcinogenesis, analysing 27,578 CpGs in each of 1,475 samples, ranging from normal cells in advance of non-invasive neoplastic transformation to non-invasive and invasive cancers and metastatic tissue. We demonstrate that hypermethylation at stem cell PolyComb Group Target genes (PCGTs) occurs in cytologically normal cells three years in advance of the first morphological neoplastic changes, while hypomethylation occurs preferentially at CpGs which are heavily methylated in Embryonic Stem Cells (MESC) and increases significantly with cancer invasion in both the epithelial and stromal tumour compartments. In contrast to PCGT hypermethylation, MESC hypomethylation progresses significantly from primary to metastatic cancer and defines a poor prognostic signature in four different gynaecological cancers. Finally, we associate expression of TET enzymes, which are involved in active DNA demethylation, to MESC hypomethylation in cancer. These findings have major implications for cancer and embryonic stem cell biology and establish the importance of systemic DNA hypomethylation for predicting prognosis in a wide range of different cancers.

**Citation:** Zhuang J, Jones A, Lee S-H, Ng E, Fiegl H, et al. (2012) The Dynamics and Prognostic Potential of DNA Methylation Changes at Stem Cell Gene Loci in Women's Cancer. *PLoS Genet* 8(2): e1002517. doi:10.1371/journal.pgen.1002517

**Editor:** Devin Absher, HudsonAlpha Institute for Biotechnology, United States of America

**Received:** October 14, 2011; **Accepted:** December 15, 2011; **Published:** February 9, 2012

**Copyright:** © 2012 Zhuang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was funded by the Eve Appeal ([www.eveappeal.org.uk](http://www.eveappeal.org.uk)) and a grant from the UCLH/UCL Comprehensive Biomedical Research Centre (project No. 152) and by the NIHR Health Technology Assessment Programme. The epigenetic analyses has been undertaken at UCLH/UCL, which received a proportion of its funding from the Department of Health NIHR Biomedical Research Centres (BRC) funding scheme. AET was supported by a Heller Research Fellowship. HCK receives support from the Manchester NIHR BRC. HBS received support from Helse Vest, Research Council of Norway, and The Norwegian Cancer Society, Harald Andersens legat. MZ was funded by IGA, Ministry of Health of the Czech Republic (project No. NS10566). This pan-European collaboration was supported by the European Network Translational Research in Gynaecological Oncology (ENTRIGO) of the European Society of Gynaecological Oncology (ESGO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: M.Widschwendter@ucl.ac.uk (MW); a.teschendorff@ucl.ac.uk (AET)

## Introduction

Aberrant DNA methylation is one of the most important cancer hallmarks [1], yet its precise role in carcinogenesis and clinical prognosis remains ill-defined [2]. Indeed, the dynamical changes in DNA methylation that happen during carcinogenesis, in particular those prior to morphological changes, have not yet been explored in detail. Moreover, no study has so far reported a DNA methylation signature capable of predicting prognosis across multiple human cancers, unlike gene expression and DNA copy number where such prognostic signatures have been described [3,4].

Both hyper and hypomethylation are commonly observed in cancer [1]. In contrast to hypomethylation, which seems to target

large inter-genic satellite repeat regions, hypermethylation appears to happen locally, preferentially targeting the promoters of genes. Several studies have reported that a statistically high fraction of these promoters map to stem cell PolyComb Group Target genes (PCGTs) [5,6], many of which encode transcription factors needed for differentiation, and which are normally suppressed in embryonic stem cells through a reversible mechanism mediated by the Polycomb Repressive Complex (PRC2) [7]. This preferential hypermethylation at PCGTs in cancer supports the view that the reversible gene repression of PCGTs in stem cells may be replaced by permanent silencing in cancer, potentially impairing the differentiation capacity of cells [1,5,6]. Although there is no causal functional data linking PCGT methylation to carcinogenesis yet, there is accumulating evidence that factors which lead to

## Author Summary

DNA methylation is an important chemical modification of DNA that can affect and regulate the activity of genes in human tissue. Abnormal DNA methylation and its subsequent effects on gene activity are a hallmark of cancer, yet when precisely these DNA methylation changes occur and how they contribute to the development of cancer remains largely unexplored. In this work we measure the methylation state of DNA at over 14,000 genes in over 1,475 samples, including normal and benign cells, invasive cancers, and metastatic cancer tissue. Using cervical cancer as a model, we show that gain of abnormal methylation at genes typically un-methylated in stem cells can be detected up to 3 years in advance of the appearance of pre-cancerous cells, while those genes typically methylated in stem cells lose this methylation progressively throughout cancer development. Furthermore, we discover that this process of methylation loss during cancer progression is a marker of poor disease outcome common to all four major women-specific cancers: breast, ovarian, endometrial, and cervical cancers. Finally we demonstrate the relationship between loss of methylation and cancer-specific over-production of a specific protein known to play an active role in removing methylation from DNA. Taken together these findings highlight the complex nature of DNA methylation dynamics in cancer development as well as their potential exploitation for clinical gain.

cancer, for instance age or oxidative damage, are causally involved in DNA methylation at PCGTs [8–11].

Another feature of the epigenetic landscape characterising human embryonic stem cells (hESC) was described by Lister et al [12]. Specifically, using single-base-resolution DNA methylation maps, they demonstrated that a substantial fraction of CpGs is heavily (>80%) Methylated in human Embryonic Stem Cells (MESC) (see Materials and Methods for the precise definition of MESC CpGs and Table S1 for the complete list of MESC CpGs on the 27 k array). However, it is unknown at present what role MESC may play throughout carcinogenesis. Thus, which epigenetic stem cell features are retained or changed in human cancer and even more importantly at which stage during human carcinogenesis these epigenetic changes occur, is still unclear.

Motivated by these outstanding questions, we decided to (i) explore the dynamics of epigenetic changes at stem cell loci (PCGTs and MESC) throughout all stages of human carcinogenesis and (ii) to investigate their potential role in predicting poor prognosis.

To address our first aim, we used as a model the uterine cervix, since screening programs in place allow easy access to this organ, and cervical carcinogenesis is also one of the few scenarios in humans where DNA methylation changes in the actual cell of origin and occurring throughout disease progression can be analyzed. Specifically, we measured DNA methylation at over 27,000 CpGs in cervical cells and at three different stages: (a) three years before onset of dysplastic changes, (b) at the stage of non-invasive dysplasia, and (c) at the stage of invasive cervical cancer. To address our second aim we analysed DNA methylation data from 5 independent cohorts encompassing a total of 1,026 tumour samples in 4 different gynaecological cancers. In total, we analysed DNA methylation data from 10 independent studies, encompassing normal and cancer tissue from 5 different tissue types, including metastases (Table 1).

Using these data we here report four major novel aspects of cancer epigenetics: (i) Hypermethylation at PCGT stem cell loci

occurs up to three years before the first signs of morphological transformation, (ii) hypomethylation at MESC stem cell loci is a hallmark of cancer invasion, affecting both epithelial and stromal compartments, and increases further in metastases, (iii) hypomethylation instability at MESC defines a stem cell DNA methylation signature that predicts poor prognosis in multiple human cancers independently of standard prognostic factors, and (iv) expression of TET enzymes [13–17] is strongly associated with MESC hypomethylation.

## Results

All methylation data in this study were generated with the Illumina Infinium Human Methylation27 beadchip array (Materials and Methods), which assesses the DNA methylation status of 27,578 CpG sites located in the promoter regions of 14,495 genes as described previously [18]. Among these CpGs, 3,465 map to PCGTs, whilst 5,943 map to MESC CpGs (Materials and Methods, Table S1 and Table S2). We also made a distinction between CpGs located within Partially Methylated Domains (PMDs) (a total of 4,750 CpGs on the array mapped to PMDs), and those that are not (termed non-PMDs). PMDs demonstrate reduced methylation levels in more differentiated embryonic tissue compared to embryonic stem cells, and consist of focally hypermethylated elements (corresponding overwhelmingly to CpG islands), concentrated within regions of long-range hypomethylation [12]. PMDs were recently described also in cancer [19]. For precise definitions see Text S1.

To investigate the dynamics of DNA methylation in human carcinogenesis we designed a study with samples from three different phases reflecting cervical carcinogenesis: (1) ‘Before Dysplasia (BDy)’: normal cervical epithelial cells collected within the ARTISTIC trial [20,21] (n = 152) of which 75 developed a cervical intraepithelial neoplasia grade 2 or 3 (CIN2/3) after three years (cases), whereas the other 77 remained normal (controls). These samples were matched for age and HPV status. (2) ‘Dysplasia (Dy)’: age-matched non-invasive dysplastic epithelial cells (CIN2/3) (n = 18, all HPV+) and normal cervical epithelial cells (n = 30, 19 HPV– and 11 HPV+) collected within screening programs [22], and (3) ‘Invasive Cancer (CA)’: invasive cervical cancer tissue (n = 48) and normal cervical tissue (n = 15) collected within a clinical setting. Further details of the samples are described in Text S1 (see also Table 1).

As expected, PCGTs were highly enriched among CpGs hypermethylated in invasive cervical cancer (Figure 1A and 1C). In contrast, CpGs that become hypomethylated in invasive cervical cancer are to a large extent MESC (Figure 1B and 1D). Most importantly, PCGTs were hypermethylated three years prior to any cytological changes (Figure 1C, OR = 2.44; 95%CI = 2.27–2.63;  $p < 10^{-100}$ ), especially for those PCGT CpGs located within PMDs (OR = 4.81; 95%CI = 4.19–5.52;  $p < 10^{-100}$ ). We verified that PCGT enrichment was also independent of HPV status ( $P < 0.005$  for HPV+ and HPV–). Notably, the frequency of hypermethylation remained fairly constant throughout the phases from non-invasive dysplasia to invasive cancer (Figure 2A and Figures S1, S2, S3, S4).

In contrast to PCGT methylation, MESC hypomethylation appears as a progressive process towards invasive cancer: whereas we observed a substantial enrichment of MESC in the normal samples three years prior to the dysplastic changes (OR = 5.69 and 9.55 for PMD and nonPMD respectively), non-invasive dysplastic samples had an increased MESC enrichment in hypomethylated CpGs (OR = 7.62 and 12.30 for PMD and nonPMD, respectively) and eventually MESC CpGs contributed most significantly to

**Table 1.** Data sets used in this study.

Data set	Sample size	Normal	Cancer	Cell type	Age range (yr)	Reference
BDy	152	152	0	LBC	19–55	GSE30760
Dy	48	30	18	LBC	26–43	GSE20080
Cvx CA	63	15	48	tissue	24–91	GSE30760
BC	60	37	23	tissue	19–75	GSE32393
BC (JHU)	118	15	103	tissue	20–96	GSE31979
EC	87	64	23	tissue	32–90	Salvesen, H.B. et al.
EC (Meta)	17	0	17	tissue	43–93	GSE33422
OvC	177	0	177	tissue	24–88	Teschendorff, A.E. et al.
OvC (TCGA)	378	0	378	tissue	34–89	TCGA
ColC	154	29	125	tissue	NA	GSE25062
LC	151	24	127	tissue	NA	TCGA
LC (Fibro)	10	5	5	fibroblast	58–77	GSE22874
<i>BRCA1</i>	60	60	0	WBC	50–80	GSE32396

Cervical cells collected three years before half of the patients developed dysplasia (BDy), normal versus dysplastic cervical cells (Dy), normal cervical tissue and invasive cervical cancer (Cvx CA), non-neoplastic breast tissue and breast cancer (BC), normal endometrium and endometrial cancer (EC) and metastatic endometrial cancer (EC (Meta)), ovarian cancer tissue (OvC; The Cancer Genome Atlas, TCGA), colon cancer tissue (ColC), lung cancer tissue (LC) and normal and cancer associated lung fibroblasts (LC Fibro), and white blood cell (WBC) samples from *BRCA1* mutation carriers and controls (*BRCA1*). The number of samples, their distribution in terms of normal and cancer, cell-type, age-range (years) and reference to data access is given. GSE numbers are GEO accessions. NA, not available.

doi:10.1371/journal.pgen.1002517.t001

hypomethylated CpGs in invasive cancer (OR = 18.84 and 26.85 for PMD and nonPMD respectively; Figure 1D, Figure 2A, and Figures S1, S2, S3, S4). In order to check that these enrichments are not just a consequence of the baseline methylation levels (i.e. the levels in normal tissue), we estimated the enrichment relative to other CpGs with specific baseline methylation levels (CpGs with mean  $\beta$ -values in normal cervical tissue samples of  $<0.2$  and  $>0.4$ ). This confirmed that the observed PCGT and MESC enrichment was independent of the initial methylation levels in normal tissue, and that this was particularly true for PCGT/MESC CpGs within PMDs (Figure S5). Thus, MESC CpGs that showed reduced methylation levels ( $<80\%$ ) in normal tissue compared to their levels in hESCs ( $>80\%$ ) were still more likely to exhibit further hypomethylation in dysplasia and cancer than a control set of CpGs with similar methylation levels in normal tissue (Figure S5).

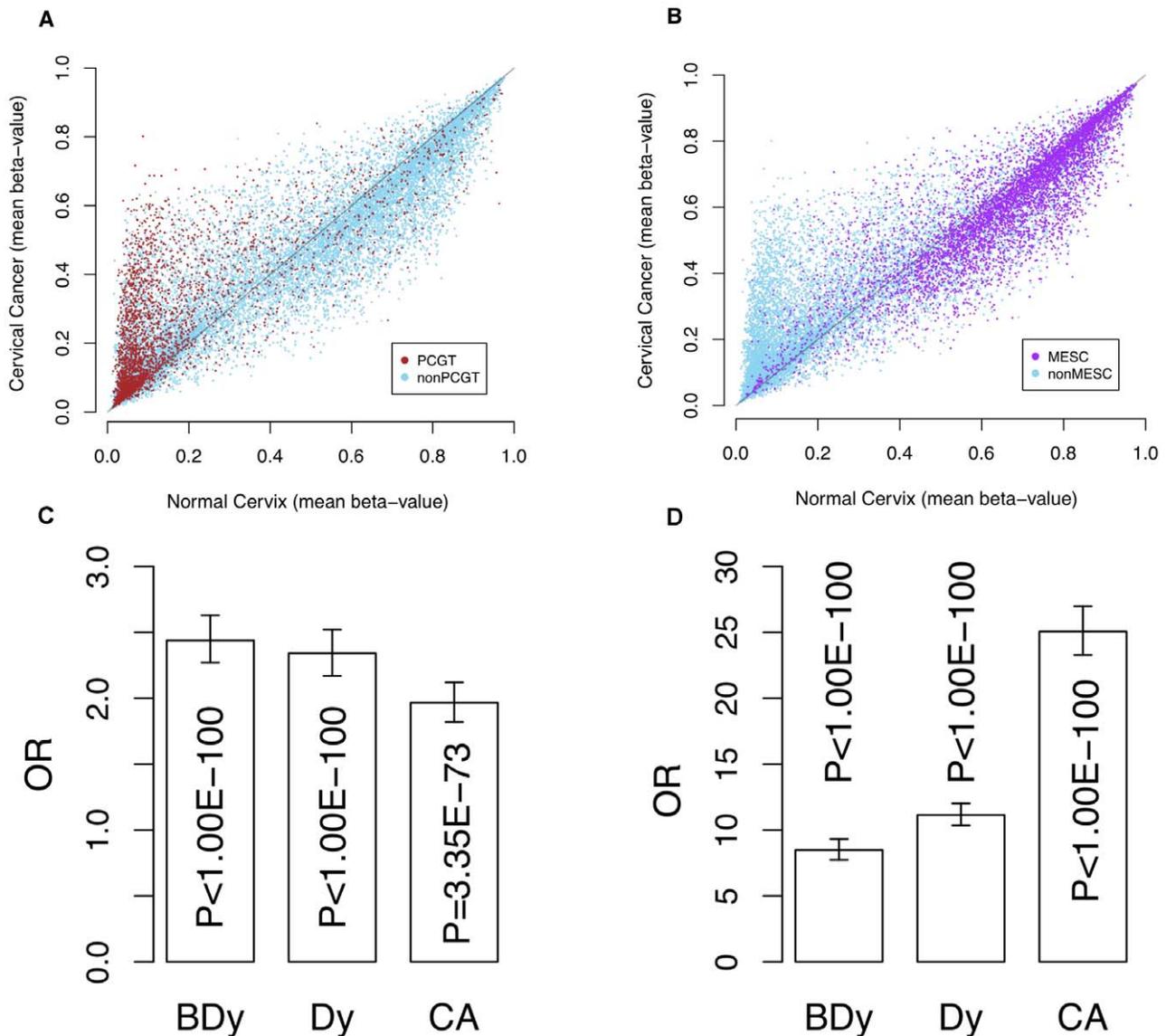
To test if PCGT and MESC methylation changes are also present in cells which are not immediately involved in carcinogenesis we studied white blood cell DNA from women who carry *BRCA1* mutations and who are therefore at an 80% lifetime risk of developing breast and/or ovarian cancer. Whereas MESC methylation was not altered, we observed that PCGTs were highly enriched among CpGs hypermethylated in blood cells from *BRCA1* mutation carriers (Figures S6 and S7).

Next, we asked if the progressive hypomethylation of MESC towards invasive cancer is a generic feature of tumour biology. We analysed DNA methylation profiles of breast, endometrial, colorectal and lung cancer (Text S1; Figure 2B and Figures S1, S2, S6, S7), and in all cancer types we observed a significant loss of methylation at MESC CpGs, concurrent with the expected hypermethylation of PCGT CpGs.

As demonstrated in Figure 2A and 2B, PCGT methylation enrichment exists prior to and at the stage of non-invasive dysplasia when analyzing only epithelial cells without stroma and remains constant when studying invasive cancer tissue which contains some stromal components. In contrast, MESC enrichment doubles in the hypomethylated fraction when comparing

invasive cancer to non-invasive dysplastic cells. This pronounced enrichment could be contributed by MESC hypomethylation in the cancer-associated stromal component. To test this, we analyzed those PCGTs and MESC CpGs that are enriched in the hyper- and hypomethylated fractions in lung cancer and asked if these CpGs are also enriched in lung cancer associated fibroblasts compared to normal lung fibroblasts [23]. Interestingly, while there was no enrichment of PCGTs (Figure 2C), there was a clear enrichment of lung cancer MESC CpGs among PMD CpGs that are hypomethylated in lung cancer fibroblasts (Figure 2D). This further supports the view that MESC hypomethylation is an important characteristic of cancer invasion, and that it may therefore be a molecular determinant of clinical outcome.

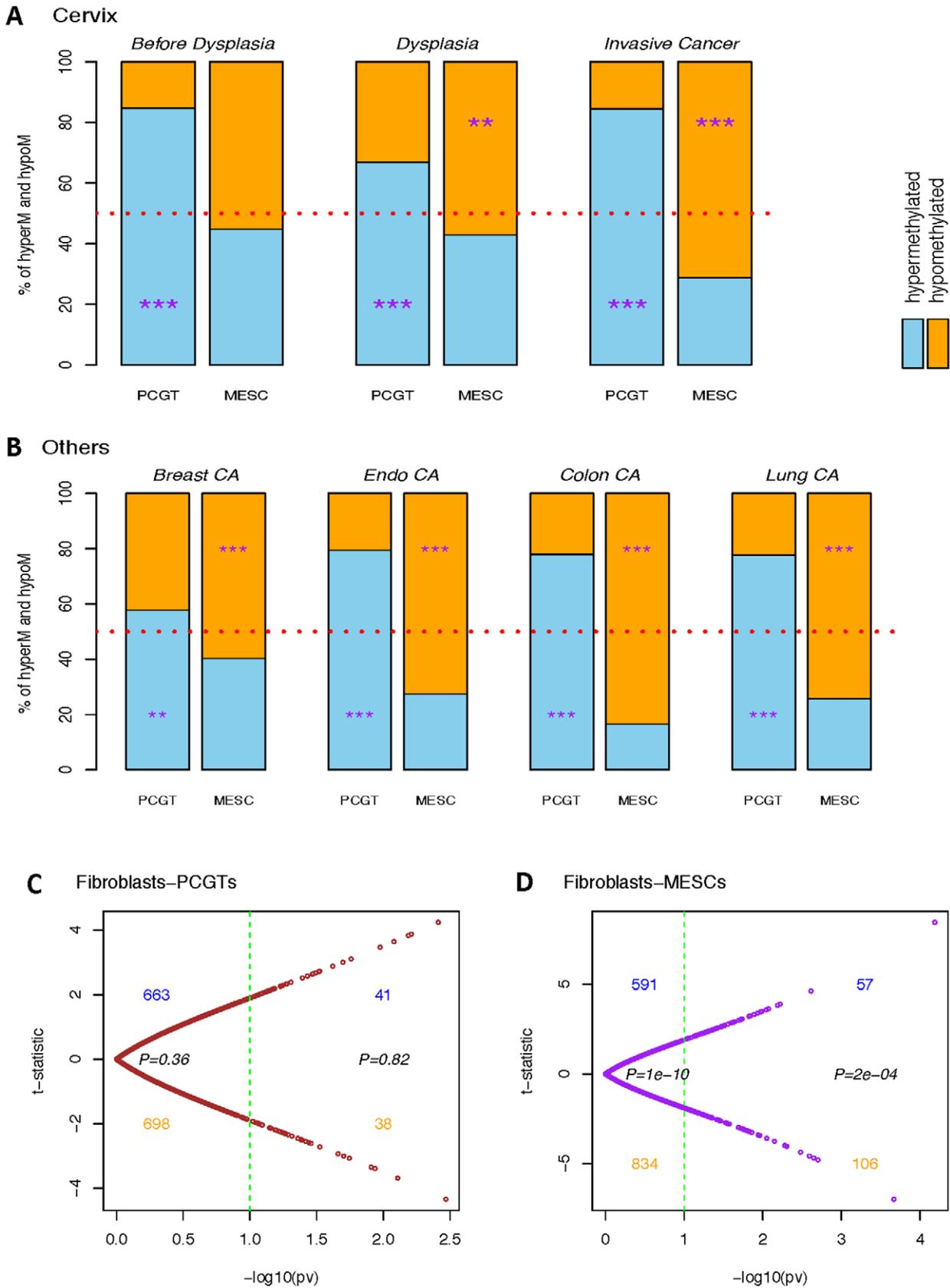
Molecular signatures, and in particular gene expression signatures, involving stem cell genes have been associated with poor prognosis in several cancers [24,25]. Therefore, given the fundamental role of PCGT and MESC CpGs in the dynamics of DNA methylation in human cancer, as just described, it is natural to ask if DNA methylation changes at these stem cell loci can predict clinical outcome. In particular, we posited that epigenetic instability, as measured by DNA methylation changes from a normal reference, might indicate clinical outcome. To test this idea, we devised an Epigenetic Instability Index (EpI) to evaluate instability for each tumour sample as the fraction of significant DNA methylation changes relative to a corresponding normal reference profile (Materials and Methods). The instability index was divided into 4 types according to the baseline normal reference methylation (0 = unmethylated, 1 = hemimethylated, 2 = methylated) and the nature of DNA methylation changes (0 $\rightarrow$ 1/2, 1 $\rightarrow$ 2, 1 $\rightarrow$ 0, 2 $\rightarrow$ 0/1) observed in cancer (Materials and Methods, Figure 3A and 3B). In addition, we considered the EpI restricted to PCGT and MESC stem cell loci, and since very few PCGT CpGs were observed to be methylated (1 or 2) in normal tissue, this resulted in 3 stem cell EpI indices: PCGT (0 $\rightarrow$ 1/2), MESC (1 $\rightarrow$ 0), MESC (2 $\rightarrow$ 0/1). Remarkably, we observed that the demethylation instability index (DeMI) at MESC CpGs (2 $\rightarrow$ 0/1) was associated with poor prognosis in endometrial, breast, ovarian,



**Figure 1. Methylation profile of PCGTs and MESC in cervix data.** Scatter plots of mean  $\beta$ -values in normal cervical tissue (x-axes) vs. cervical cancer tissue (y-axes) of (A) all CpGs with PCGTs highlighted in brown and (B) all CpGs with MESC highlighted in purple. (C) is the bar chart indicating the enrichment odds ratios (OR) and P-values (Fisher-test), testing for enrichment of PCGTs among CpGs unmethylated in normal cervix (mean  $\beta$ -value  $< 0.2$ ) and with a higher mean  $\beta$ -value in (i) normal samples which develop dysplasia (BDy), (ii) non-invasive dysplastic samples (Dy), and (iii) invasive cervical cancer (CA); (D) the bar chart indicating the enrichment of MESC among those CpGs methylated in the normal cervix (mean  $\beta$ -value  $> 0.4$ ) and with a lower mean  $\beta$ -value in tissue representing the three stages of cervical carcinogenesis (BDy, Dy, and CA). doi:10.1371/journal.pgen.1002517.g001

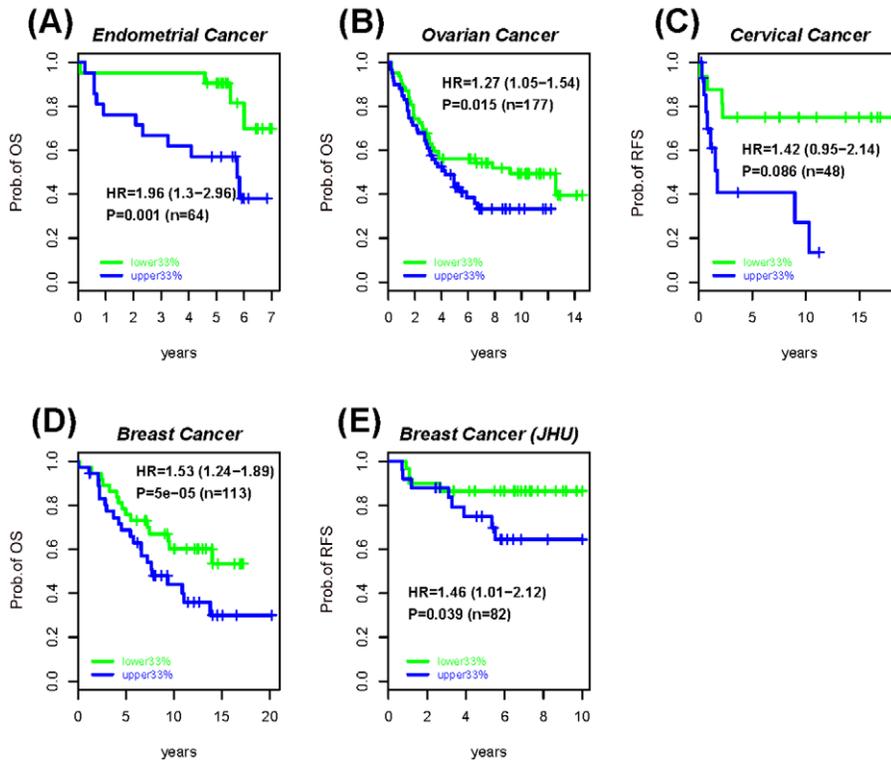
and cervical cancers (Figure 4). In multivariate analysis, the DeMI was a predictor of poor prognosis in all cancers independently of other prognostic factors (Table 2 and Table S3), demonstrating the clinical potential of this DNA methylation stem cell signature. In contrast, the methylation instability index defined at PCGTs only correlated with clinical outcome in ovarian cancer (Table S3). Survival analysis at individual CpG level further demonstrated the consistent enrichment of MESC CpGs among prognostic CpGs hypomethylated in poor outcome samples in all 4 invasive cancers, whereas PCGT CpGs were not consistently enriched in either the hyper or hypomethylated prognostic component (Table S4). There was also substantial overlap between the MESC CpGs which have stable methylation levels in normal tissue and which become hypomethylated in cancer, and prognostic MESC CpGs that are hypomethylated in poor outcome tumour samples (Table S5).

To further demonstrate that MESC hypomethylation is an important determinant of poor outcome in human cancer, we tested if these epigenetic changes progress further in metastatic cancer. Thus, we compared DNA methylation profiles of primary endometrial cancers to extra-uterine metastases of endometrial cancer. Importantly, the DeMI index was higher in metastatic cancer compared to primary tumours, but not so for the hypermethylation instability index at PCGTs (Figure 5A). In fact, the DeMI index demonstrates clinical potential for discriminating primaries that may be destined to metastasize (Figure 5B). From these data we can therefore conclude that while PCGT hypermethylation is an important event in early oncogenesis, which persists at later stages, MESC hypomethylation is a progressive process and a key characteristic of more malignant cancers (Figure 3B).



**Figure 2. Differential dynamics of hypermethylated and hypomethylated PCGTs and MESCs.** Bar charts representing percentages of significantly hypermethylated (blue) and hypomethylated (orange) PCGT and MESC CpGs in (A) each stage of cervical carcinogenesis: Cervix 'Before Dysplasia', 'Dysplasia', and 'Invasive Cancer', all relative to normal cervical cells or tissue; and in (B) 'Breast CA', 'Endo CA', 'Colon CA', and 'Lung CA', all





**Figure 4. Survival analysis of the MESC demethylation instability index in various cancers.** Kaplan-Meier survival curves between the upper (blue) and lower (green) tertiles of the demethylation instability index (DeMI) at MESC in (A) Endometrial cancer, (B) Ovarian cancer, (C) Cervical cancer, and (D–E) two Breast Cancer cohorts. The hazard ratio (HR), 95% confidence interval (CI) and P-values are from the multivariate Cox regression model, with “n” denoting the number of samples in cohort. Clinical endpoint used is indicated on the y-axis (OS=overall survival, RFS=relapse free survival).

doi:10.1371/journal.pgen.1002517.g004

hydroxylation is involved in epigenetic reprogramming of the zygotic paternal DNA following natural fertilization and that this may also contribute to somatic cell nuclear reprogramming during animal cloning [16]. We therefore analysed mRNA expression of TET1 and two isoforms of TET2, and TET3 (see Text S1 for primer information), to test whether hypomethylation is associated

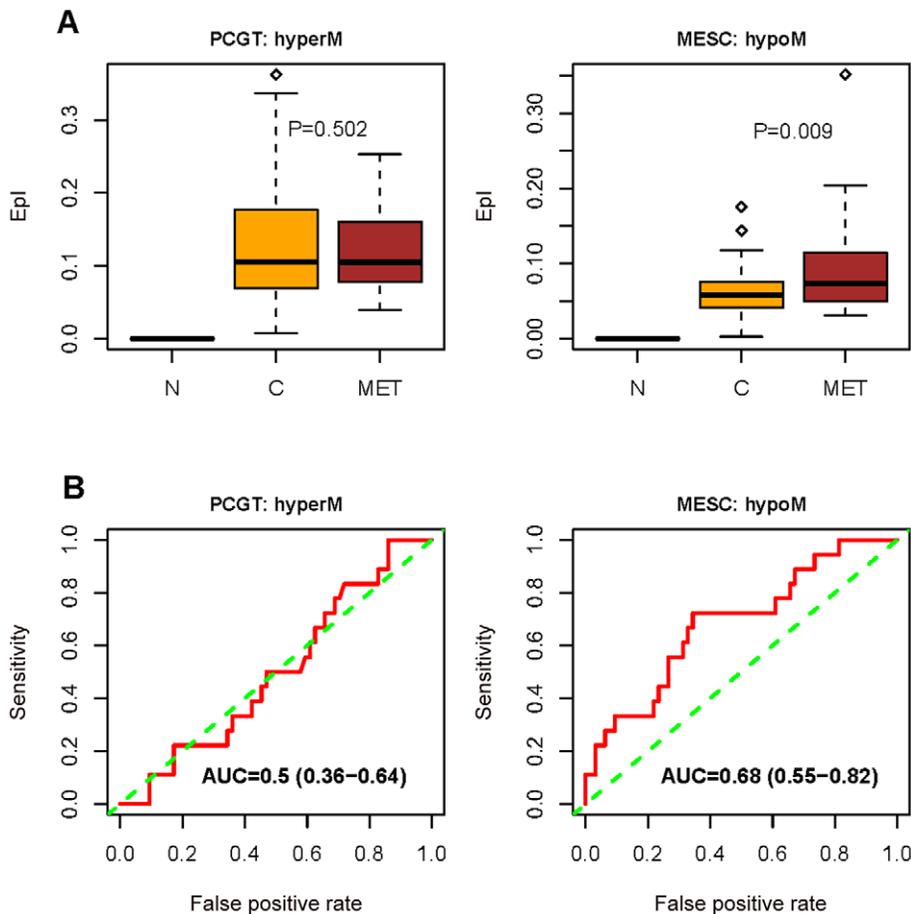
with TET expression. We observed a strong correlation between high TET, in particular TET3 expression, and hypomethylation, specifically at MESC CpGs (Figure 7 and Figure S8). We checked that the anti-correlation of TET expression with MESC CpG methylation was independent of the level of methylation in normal tissue (Figure S9). Although this observation is purely correlative, it

**Table 2.** Survival analysis results of the PCGTs and MESC Epigenetic Instability Index (Epi) in women’s cancers.

PCGT Epi	Univariate			Multivariate		
	HR (95%CI)	P	n	HR (95%CI)	P	n
Endo CA	0.80 (0.48–1.33)	0.385	64	0.90 (0.53–1.54)	0.71	63
Ovarian CA	1.34 (1.13–1.59)	7.00E-04	177	1.54 (1.24–1.92)	9.00E-05	161
Cervical CA	0.75 (0.48–1.19)	0.222	48	0.96 (0.59–1.56)	0.878	47
Breast CA	0.99 (0.76–1.29)	0.949	113	0.93 (0.70–1.25)	0.645	107
MESC Epi (DeMI)	HR (95%CI)	P	n	HR (95%CI)	P	n
Endo CA	1.96 (1.30–2.96)	0.001	64	2.18 (1.24–3.83)	0.007	63
Ovarian CA	1.27 (1.05–1.54)	0.015	177	1.40 (1.09–1.79)	0.008	161
Cervical CA	1.42 (0.95–2.14)	0.086	48	1.47 (0.95–2.28)	0.085	47
Breast CA	1.53 (1.24–1.89)	5.00E-05	113	1.51 (1.20–1.89)	5.00E-04	107

Univariate and multivariate Cox regression results for the PCGT Epi and MESC Epi (DeMI) in endometrial, ovarian, cervical and breast cancer with number of samples (n), Hazard ratio (HR), 95% confidence interval (CI), and P-value (P). Overall survival was used for endometrial, ovarian and breast cancer, relapse free survival for cervical cancer.

doi:10.1371/journal.pgen.1002517.t002



**Figure 5. Methylation changes between normal, primary, and metastatic endometrial cancer.** (A) Boxplots comparing the frequency of PCGT (0→1/2) DNA methylation changes, and the frequency of combined MESC (1→0) and MESC (2→0/1) DNA methylation changes (“combined DeMI index”), between normal endometrium (N), primary endometrial cancer (C), and between normal endometrium and metastatic endometrial cancers (MET). One-tailed Wilcoxon rank sum test P-values for the instability indices between cancer and metastases are indicated. (B) Receiver operating characteristic (ROC) curves measuring the dissimilarity in the combined DeMI index between primary and metastatic endometrial cancers with corresponding Area Under Curve (AUC) and 95% CI. doi:10.1371/journal.pgen.1002517.g005

is consistent with the view that TET3 overexpression (Figure S10) in cancer contributes to reprogramming of cancer cells via active DNA demethylation.

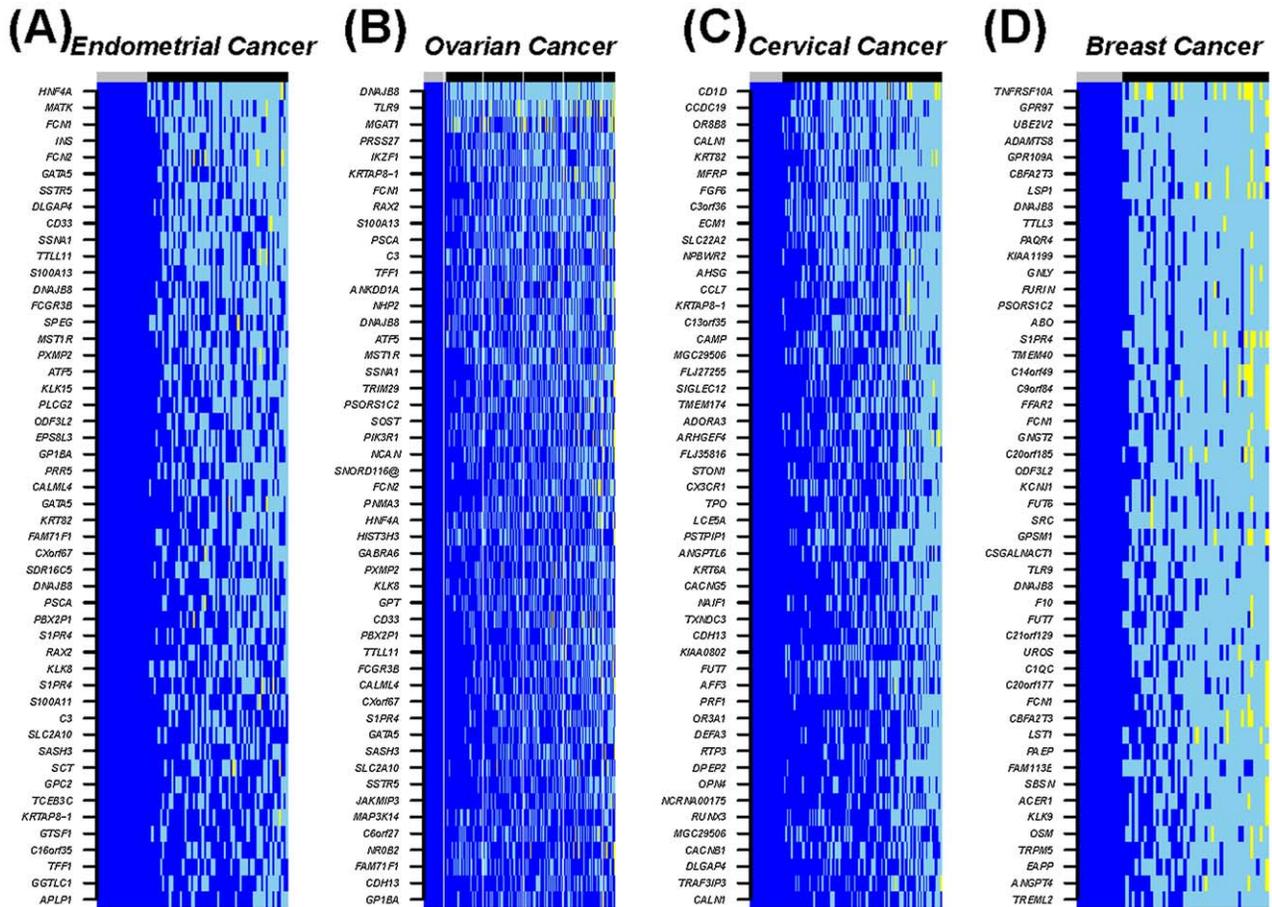
## Discussion

Epithelial cells of the uterine cervix offer a unique opportunity to study epigenetic alterations throughout carcinogenesis. Our first key result is the demonstration that normal cells of origin acquire methylation changes at least three years in advance of the first morphological changes. Specifically, our data demonstrate that PCGT hypermethylation and MESC hypomethylation are major contributors to early cervical carcinogenesis. This is independent of human papillomavirus (HPV) infection as our study was matched for HPV status, and since PCGT enrichment was observed in both HPV+ and HPV- samples. Importantly, the observed enrichments were also independent of the levels of methylation in normal tissue. That is, MESC which showed full methylation (i.e.  $\beta$ -value>0.8) or hemi-methylation (i.e.  $0.3 < \beta$ -value<0.7) were preferentially hypomethylated in dysplasia and cancer in comparison to control sets of CpGs with same methylation levels in normal tissue.

The role of PCGT methylation as a very early event is further supported by our finding that PCGTs were highly enriched among CpGs which were hypermethylated in blood cells from *BRCA1* mutation carriers, suggesting that *BRCA1* is an important regulator of the DNA methylome and that aberrant *BRCA1* function could lead to increased predisposition to cancer through increased methylation at PCGT loci. The fact that *BRCA1* mutation carriers showed increased PCGT methylation in their blood cells but are at no substantial increased risk to develop blood-borne cancers suggests that PCGT hypermethylation refers a substantial risk but that there are additional factors required (e.g. endocrine, paracrine or viral triggers).

Our second key result is that MESC hypomethylation occurs in both the epithelial and stromal components of cancer and that this is a progressive process, increasing significantly towards invasion and metastatic cancer. This in turn suggests that the level of MESC hypomethylation in primary tumours may be an important determinant of clinical outcome.

Indeed, our third key result is the report of a stem cell (MESC) DNA hypomethylation signature that can predict clinical outcome in multiple human cancers, independently of known prognostic factors. To the best of our knowledge this constitutes the first



**Figure 6. Heatmap of the top 50 most frequently hypomethylated MESC CpGs in cancers.** The CpGs show stable fully methylated states across all normal samples. Methylation values: blue=high methylation ( $\beta$ -value > 0.7), skyblue= hemi methylation ( $0.25 < \beta$ -value < 0.7), yellow = low methylation ( $\beta$ -value < 0.25). Sample labels at the top of the heatmaps: normal (grey) and cancer (black). doi:10.1371/journal.pgen.1002517.g006

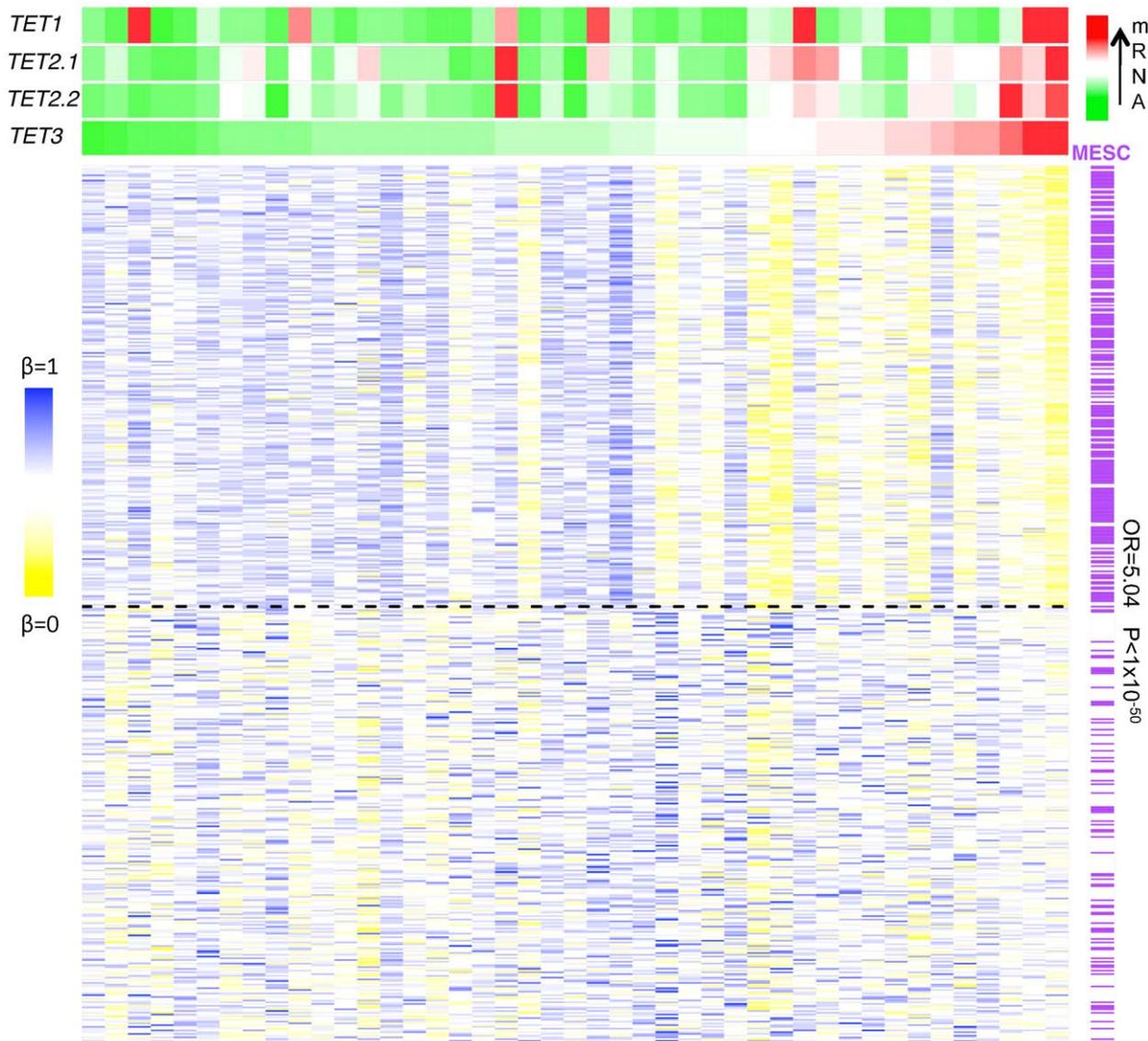
report of a common prognostic signature in cancer that is based on DNA methylation, and is therefore an epigenetic analogue to the prognostic genomic instability signature presented in [3].

Besides the key distinction of PCGT and MESC CpGs, we also observed that the localisation of CpGs in relation to PMDs was another important facet of the pattern of DNA methylation changes. Specifically, PCGT hypermethylation was observed preferentially within PMDs, while the progressive MESC hypomethylation in cancer was equally strong in PMDs and non-PMDs. We point out that while the PMDs considered here were defined for colon cancer cells, that these broad regions of partial methylation overlap significantly between colon tissue and fibroblasts, suggesting that these regions may be largely similar also between different tissues.

The similarities between normal developmental and cancer epigenetic programming are intriguing. While embryonic stem cells suppress differentiation-inducing genes reversibly via promoter occupancy of PRC2, cancer cells suppress these same genes much more robustly via covalent DNA modification. Even more interestingly, trophoblast cells whose core function is to invade the maternal tissue and form the placenta, are relatively more hypomethylated compared with the inner cell mass, which will differentiate into the embryo [28], supporting the view that hypomethylation may be associated with the capacity to invade neighbouring tissue such as the maternal endometrium. Similarly,

the observed correlation between MESC hypomethylation and the malignant potential of cancers suggests that fully methylated MESC CpGs may provide a protective mechanism against invasion. Thus, the fact that the great majority of MESC CpGs exhibit similar high methylation levels in stem cells and normal tissues, means that high MESC methylation may be viewed as an intrinsic property of any normal cell, regardless of whether it is a stem cell or a mature differentiated one. In this model then, hypomethylation at MESC CpGs would lead to a transformed cellular phenotype that is more prone to invasion. In this context however, it is worth pointing out that the observed MESC hypomethylation could also be reflecting changes in the stromal cell content of the tumours. Indeed, the observation that cancer fibroblasts show similar hypomethylation changes at MESC loci suggests that the more frequent MESC hypomethylation in invasive cancers could be partly due to increased numbers of cancer fibroblasts.

It could also be argued that the other DNA methylation changes we have reported here are the result of changes in the stromal and immune cell compartments of the tumours. However, we verified using Principal Components Analysis (PCA) and GSEA analysis [26] on normal liquid based cytology (LBC) samples and separately on age-matched cervical dysplasias (Table 1, ‘‘Dy’’-study) that the components of variation associated with stromal and immune cell markers were very similar between normal and dysplasia, in stark contrast to PCGTs which showed a dramatic



**Figure 7. Anti-correlation between the TET mRNA expression level and methylation  $\beta$ -value.** Heatmap of the 250 most (upper half) and least (lower half) anti-correlated hypomethylated CpGs in cervical cancer samples ranked according to their TET3 mRNA expression levels from the lowest (green) to the highest (red) with the indication of MESC (purple) and nonMESC (white). The odds ratio (OR) and P-value (P) are obtained from Fisher's exact test estimating the enrichment of MESC among hypomethylated CpGs that are significantly anti-correlated with TET3 mRNA expression level.

doi:10.1371/journal.pgen.1002517.g007

difference with comparatively no variation in normal tissue but representing the dominant component of variation in dysplasia (manuscript in preparation). Thus, the DNA methylation changes at PCGT loci reported here are unlikely to be due to changes in the stromal cell composition of tumours.

Finally, the crucial role of TET3 in DNA demethylation and early development, its overexpression in cancer, and the observed correlation with MESC hypomethylation, supports the view that aberrant developmental programs leading to reprogramming of the epigenome in adult cells may be critical for carcinogenesis. Interfering with these aberrant programs may therefore lead to novel ways to treat cancer.

In summary, our findings suggest that epigenetic deregulation of two distinct sets of genes, both important for stem cell integrity, impact carcinogenesis in different ways: one process involves gain

of methylation and is a hallmark of de-differentiation and early oncogenesis, while the other involves loss of methylation and is a key determinant of invasion and clinical outcome.

## Materials and Methods

### Definition of MESC

A recent study used bisulfite sequencing to map, at single-base-resolution, DNA methylation throughout the majority of the human genome in both embryonic stem cells and fibroblasts [12]. For each CpG site, the number of C and T reads covering each methyl cytosine on both forward and reverse strands were provided [12]. The multiple reads covering each methyl cytosine can be used as readout of the fraction of sequences within the sample that are methylated at that particular site (i.e. C reads/

C+T reads) [29], and hence, referred as the methylation level of the site. In this study, Methylated in human Embryonic Stem Cells (MESC) CpGs are the CpG sites that were covered by at least 5 reads on both forward and reverse strands (i.e. the total number of C and T reads on both strands  $>=5$ ) and the overall mean methylation levels (i.e. the average methylation level of both the forward and reverse strands) is greater than 80%. MESC CpGs were then mapped to those present on the Illumina 27 k array (Table S1). Functional annotation (gene assignment) of the MESC CpGs present on the array was obtained from Illumina and Bioconductor annotation packages.

### Definition of PCGTs

PolyComb Group Target genes (PCGTs) were defined as CpGs which are occupied by SUZ12 and/or Eed and/or are trimethylated at Lysine 27 on histone H3 in human embryonic stem cells (Table S2, annotation file kindly provided by Benjamin P. Berman and Peter W. Laird) [19].

### DNA Methylation Assay

**DNA extraction.** DNA from LBC samples and tissues was isolated using the Qiagen DNeasy Blood and Tissue Kit (Qiagen Ltd, UK, 69506) and quantified via spectrophotometry (Nanodrop, Thermo Scientific Ltd UK) with 600 ng DNA from each sample. DNA from whole blood was extracted using a chloroform based extraction method from 400  $\mu$ L of blood. All DNA samples were bisulphite modified using the EZ DNA Methylation Kit D5004/8 (Zymo Research, Orange, CA, USA) according to the manufacturer's instructions.

**DNA methylation profiling.** The genome wide methylation analyses were performed using the validated Illumina Infinium Human Methylation27 BeadChip (Illumina Inc USA, WG-311-1201) [18]. During the assay, bisulphite (BS) converted DNA is amplified, fragmented and hybridised to the BeadChip arrays (each chip accommodates 12 samples as designated by Sentrix positions A–L). A single base extension is then performed using labelled DNP- and biotin labelled dNTPs. The arrays were imaged using a BeadArray Reader. Image processing and intensity data extraction were performed according to Illumina's instructions. Each interrogated locus is represented by specific oligomers linked to two bead types: one representing the sequence for methylated DNA (M) and the other for unmethylated DNA (U). For each specific CpG site, the methylation status is calculated from the intensity of the M and U alleles, as the ratio of the fluorescent signals  $\beta = \text{Max}(M,0) / [\text{Max}(M,0) + \text{Max}(U,0) + 100]$ . Hence, DNA methylation  $\beta$ -values are continuous variables between 0 (absent methylation) and 1 (completely methylated) representing the ratio of the methylated allele to the combined locus intensity.

**TET expression.** Total RNA was isolated as previously described [30]. Reverse transcription of RNA was performed using M-MLV Reverse Transcriptase (Promega) according to the manufacturer's instructions. Primers and probes for the TET genes were designed using Primer Express (Applied Biosystems, Foster City, CA, USA). Samples in which TET was not amplified by real-time PCR after 45 cycles were classified as TET negative.

### Statistical Methods

**Quality control and inter-array normalisation.** Quality control procedures and intra-array normalisation were run on all data except the 'Colon CA', 'Lung CA', and 'Ovarian CA' sets, for which the intra-array normalised data was downloaded directly from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) databases. Background corrected *U* and *M* values,

$\beta$ -values (as generated from the Beadstudio software) and built-in controls were used to evaluate the quality of individual arrays. Samples with low BS conversion efficiency (BS control intensity values  $<4000$ ) were excluded, as well as other outliers that we detected using boxplots of total intensity  $I = U + M$  values and histograms of  $\beta$ -values. Samples were filtered further according to CpG coverage, using the Beadstudio P-values of detection of signal above background.

**Enrichment analysis.** Enrichment analysis was performed using a two-tailed Fisher's exact test. Odds ratios (OR) and 95% confidence intervals (CI) of enrichment were also computed and their corresponding significance levels estimated. Enrichment analysis was performed with a range of thresholds to check for robustness and using the Infinium 27 k array as reference to avoid array-specific bias.

**Supervised analysis.** A linear regression approach was used to model the association between disease status (cases or controls) and the CpG  $\beta$ -value methylation profile. Adjustment for age and experimental factors (e.g. bisulphite conversion) was performed by inclusion of these factors in the model as covariates. Chip effects were observed, and in this study all data were adjusted by either applying the "ComBat" method (a method that is robust to outliers and that allows for adjustment in cases where sample sizes per chip are small) [31] or using the chip as a covariate in the linear model. The linear model was adopted over a non-linear logistic or probit model as the linear model performed better in capturing profiles with larger effect sizes.

**Skewness analysis.** Given the two disease-status-associated CpG lists (hyper- or hypomethylated) obtained from the supervised analysis, the two-tailed binomial test was used to detect the skewness of the methylation in various categories (i.e. colon-PMD PCGTs, colon-PMD MESC, nonPMD PCGTs, and nonPMD MESC) of the CpGs (Figure 2, Figures S1 and S2).

**Epigenetic instability analysis.** We devised an Epigenetic Instability Index (EpI) for each tumour sample as follows. First, CpG readings were defined as unmethylated (0) ( $\beta$ -value  $<0.25$ ), hemimethylated (1) ( $0.25 \leq \beta$ -value  $\leq 0.7$ ), and methylated (2) ( $\beta$ -value  $>0.7$ ). Next, we selected CpGs with stable methylation profiles in normal tissue, defined as those CpGs with the same methylation state in all normal samples corresponding to the given tissue. These stable CpGs can undergo four types of DNA methylation changes in cancer:  $0 \rightarrow 1/2$ ,  $1 \rightarrow 2$ ,  $1 \rightarrow 0$  and  $2 \rightarrow 0/1$ . Therefore, for each tumour sample, we computed four different "instability" indices, reflecting the fraction of stable CpGs undergoing the specific types of DNA methylation changes shown. When computing these indices, and to ensure their robustness to the choice of methylation thresholds above, we also required at least a 10% change in  $\beta$ -values for calling DNA methylation differences between normal and cancer tissue. This buffering therefore avoids calling potentially small differences in  $\beta$ -values ( $<10\%$ ), which nevertheless may trespass the methylation thresholds (0.25, 0.7) used. The EpI indices were also computed by restricting the set of stable CpGs to those mapping to PCGT and MESC stem cell loci. Since the great majority of PCGT CpGs were observed to be stably unmethylated (0) in normal tissue, this resulted in 3 "stem cell EpI" indices: PCGT ( $0 \rightarrow 1/2$ ), MESC ( $1 \rightarrow 0$ ), MESC ( $2 \rightarrow 0/1$ ). We call the latter index the Demethylation instability index (DeMI).

**Survival analysis.** Univariate and multivariate Cox regression models were used for the survival analysis. In the multivariate analysis, besides DNA methylation  $\beta$ -values (or the EpI index), those clinical and histological factors, which were associated with survival in univariate analysis were also included as covariates.

## Supporting Information

**Figure S1** Differential dynamics of hypermethylated and hypomethylated PMD PCGTs and PMD MESC CpGs. Bar charts representing percentages of significantly hypermethylated (blue) and hypomethylated (orange) PMD PCGT and PMD MESC CpGs in (A) each stage of cervical carcinogenesis: Cervix 'Before Dysplasia', 'Dysplasia', and 'Invasive Cancer', all relative to normal cervix tissue; and in (B) 'Breast CA', 'Endo CA', 'Colon CA', and 'Lung CA', all relative to their respective normal controls. The significance of the binomial test assessing skew of hypermethylated versus hypomethylated (Figure S3 and S4) is indicated by \*, \*\*, and \*\*\* for P-value < 0.05, 0.01, and 0.001 respectively.

(TIF)

**Figure S2** Differential dynamics of hypermethylated and hypomethylated nonPMD PCGTs and nonPMD MESC CpGs. Bar charts of the percentages between the disease (or mutation) status associated hypermethylated (blue) and hypomethylated (orange) for nonPMD polycomb group target gene (PCGT) CpGs and nonPMD methylated in human embryonic stem cells (MESC) CpGs that pass their corresponding significance level thresholds (the same notation as in Figure S1).

(TIF)

**Figure S3** Statistical output from linear regression model estimating the association of the PMD PCGT and PMD MESC CpGs to the outcomes of the three stages of cervical carcinogenesis. Scatterplots of the linear regression fitted (adjusted for age, chip and bisulphite conversion) t-statistics against their corresponding  $-\log_{10}$ (P-values) that test the association with the cases and controls of the cervix 'Before Dysplasia' (CIN2/3 status), 'Dysplasia' (CIN2/3 status), and 'Invasive Cancer' (cancer status) on the PMD PCGT (left column) and PMD MESC (right column) CpGs. Green vertical lines denote the significant level thresholds of P-value = 0.1 for 'Before Dysplasia' and 'Dysplasia', and 0.001 for 'Invasive Cancer'. The overall numbers of CpGs that are hypermethylated (blue) and hypomethylated (orange) with their associated two-sided Binomial test P-value are given on the left hand side of the P-value threshold lines and the number of CpGs that are hypermethylated (blue) and hypomethylated (orange) pass the corresponding P-value threshold with their Binomial test P-value on the right.

(TIF)

**Figure S4** Statistical output from linear regression model estimating the association of the nonPMD PCGT and nonPMD MESC CpGs to the outcomes of the three stages of cervical carcinogenesis. Scatterplots of three cervical sets, similar to Figure S3, but based on the nonPMD PCGT (left column) and nonPMD MESC (right column) CpGs.

(TIF)

**Figure S5** Enrichment analysis of PMD PCGTs and PMD MESC CpGs in the hyper- and hypomethylated cervical cancer CpGs. Cumulative enrichment analysis (Fisher's exact tests ORs and P-values) of PCGTs among CpGs unmethylated (mean  $\beta$ -value < 0.2 in normal cervix) in normal cervix and which become hypermethylated in (i) normal samples three years prior to dysplasia (BDy), (ii) non-invasive dysplastic samples (Dy), and (iii) invasive cervical cancer (CA) in PMDs (A) and nonPMDs (B) respectively. Similarly, enrichment of MESC CpGs methylated (mean  $\beta$ -value > 0.4 in normal cervix) in normal cervix and that become hypomethylated in cases in PMDs (C) and nonPMDs (D) respectively.

(TIF)

**Figure S6** Statistical output from linear regression models estimating the association of the PMD PCGTs and PMD MESC CpGs to outcomes in five cohorts. Scatterplots of the linear regression fitted (adjusted for age, chip and bisulphite conversion) t-statistics against their corresponding  $-\log_{10}$ (P-values) testing the association with the cases and controls of 'Breast CA' (cancer status), 'Endo CA' (cancer status), 'Colon CA' (cancer status), 'Lung CA' (cancer status), and 'BRCA1 MUT' (BRCA1 status) on the PMD PCGT and PMD MESC CpGs. Green vertical lines denote the significant level thresholds of P-value = 0.1 for 'BRCA1 MUT' and 0.001 for all the others. The overall number of CpGs that are hypermethylated (blue) and hypomethylated (orange) with their associated two-sided Binomial test P-value are given on the left hand side of the P-value threshold lines. The number of CpGs that are hypermethylated (blue) and hypomethylated (orange) pass the corresponding P-value threshold with their Binomial test P-values on the right.

(TIF)

**Figure S7** Statistical output from linear regression models estimating the association of the nonPMD PCGT and nonPMD MESC CpGs to the outcomes in five cohorts. Scatterplots of five cohorts, similar to Figure S6, based on the nonPMD PCGT (left column) and nonPMD MESC (right column) CpGs.

(TIF)

**Figure S8** Magnitude of the anti-correlation between hypomethylated cervical cancer CpGs and TET mRNA. Hypomethylated MESC CpGs are significantly higher anti-correlated with TET2 and TET3 mRNA expression levels than the hypomethylated nonMESC CpGs in the cervical cancer samples. P-values are obtained from the Wilcoxon one-sided test.

(TIF)

**Figure S9** Magnitude of the anti-correlation between hypomethylated cervical cancer CpGs (grouped) and TET mRNA. Higher anti-correlated signature with TET2 and TET3 mRNA expression levels among hypomethylated MESC CpGs than hypomethylated nonMESC CpGs in the cervical cancer samples independent from the chosen baselines of the methylated and hemimethylated CpGs (mean  $\beta$ -value in normals > 0.4). P-values are obtained from the Wilcoxon one-sided test.

(TIF)

**Figure S10** TET mRNA expression level comparison between the normal cervix and cervical cancers. Boxplots of TET1, TET2.1, TET2.2 and TET3 mRNA expression levels of the normal cervix and cervical cancers. P-values are obtained from the Wilcoxon two-sided test.

(TIF)

**Table S1** The MESC CpGs mapped to 27 k Infinium array. List of the 5,943 MESC CpGs that mapped to the Illumina Infinium Human Methylation27 beadchip array with the information of their IlluminaID, geneID, gene symbol, MapInfo, and chromosome.

(XLS)

**Table S2** The PCGT CpGs mapped to 27 k Infinium array. Similar to Table S1, the list of the 3,465 PCGT CpGs that mapped to Illumina Infinium Human Methylation27 beadchip array.

(XLS)

**Table S3** Summary tables of the uni- and multivariate Cox regression model analysis of the PCGTs and MESC CpGs Epigenetic Instability Index (EpI) in endometrial, breast, ovarian, and cervical cancers. Univariate (UV) and multivariate (MV) Cox regression results for the PCGTs and MESC CpGs EpI in endometrial, breast, ovarian and cervical cancer overall survival (OS) and relapse free

survival (REL) with number of samples (n), Hazard ratio (HR), 95% confident interval, and P-value (P). (XLS)

**Table S4** PCGT and MESC enrichment analysis amongst hypermethylated and hypomethylated cancer prognostic CpGs. Enrichment analysis (Fisher's exact tests odds ratios (OR), 95% confidence intervals (CI) and P-values (P)) of PCGTs (colon-PMD and nonPMD) and MESC's (colon-PMD and nonPMD) among the top 500 hypermethylated and hypomethylated cancer (cervical, breast, endometrial, and ovarian cancers respectively) prognostic CpGs. (XLS)

**Table S5** Overlap between MESC CpGs with stable methylation levels in normal tissue and that become hypomethylated in cancer with the top ranked 1,000 prognostic MESC CpGs that are hypomethylated in poor outcome samples in endometrial, breast, ovarian, and cervical cancers. (XLS)

**Table S6** MESC CpGs according to the frequency of hypomethylation (cancer vs normal) as defined by the demethylation ( $2 > 0/1$ ) instability index (DeMI) in the endometrial, breast, ovarian, and cervical cancers. (XLS)

## References

- Jones PA, Baylin SB (2007) The epigenomics of cancer. *Cell* 128: 683–692.
- Baylin SB, Jones PA (2011) A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* 11: 726–734.
- Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z (2006) A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet* 38: 1043–1048.
- Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, et al. (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science* 331: 593–596.
- Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, et al. (2007) Epigenetic stem cell signature in cancer. *Nat Genet* 39: 157–158.
- Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, et al. (2007) A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* 39: 237–242.
- Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125: 301–313.
- O'Hagan HM, Wang W, Sen S, Destefano Shields C, Lee SS, et al. (2011) Oxidative Damage Targets Complexes Containing DNA Methyltransferases, SIRT1, and Polycomb Members to Promoter CpG Islands. *Cancer Cell* 20: 606–619.
- Maegawa S, Hinkal G, Kim HS, Shen L, Zhang L, et al. (2010) Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res* 20: 332–340.
- Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, et al. (2010) Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res* 20: 434–439.
- Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, et al. (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* 20: 440–446.
- Lister R PM, Downen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462: 315–322.
- Pastor WA, Pape UJ, Huang Y, Henderson HR, Lister R, et al. (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* 473: 394–397.
- Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PA, et al. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* 473: 343–348.
- Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, et al. (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* 473: 398–402.
- Gu TP, Guo F, Yang H, Wu HP, Xu GF, et al. (2011) The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature*.
- Wu H, D'Alessio AC, Ito S, Xia K, Wang Z, et al. (2011) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* 473: 389–393.
- Bibikova M, Fan JB (2010) Genome-wide DNA methylation profiling. *Wiley Interdiscip Rev Syst Biol Med* 2: 210–223.
- Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, et al. (2011) Regions of focal DNA hypermethylation and long range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature Genetics*.
- Kitchener HC, Almonte M, Gilham C, Dowie R, Stoykova B, et al. (2009) ARTISTIC: a randomised trial of human papillomavirus (HPV) testing in primary cervical screening. *Health Technol Assess* 13: 1–150, iii–iv.
- Kitchener HC, Almonte M, Thomson C, Wheeler P, Sargent A, et al. (2009) HPV testing in combination with liquid-based cytology in primary cervical screening (ARTISTIC): a randomised controlled trial. *Lancet Oncol* 10: 672–682.
- Apostolidou S, Hadwin R, Burnell M, Jones A, Baff D, et al. (2009) DNA methylation analysis in liquid-based cytology for cervical cancer screening. *Int J Cancer* 125: 2995–3002.
- Navab R, Strumpf D, Bandarchi B, Zhu CQ, Pintilie M, et al. (2011) Prognostic gene-expression signature of carcinoma-associated fibroblasts in non-small cell lung cancer. *Proc Natl Acad Sci U S A* 108: 7160–7165.
- Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, et al. (2008) An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* 40: 499–507.
- Eppert K, Takenaka K, Lechman ER, Waldron L, Nilsson B, et al. (2011) Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat Med* 17: 1086–1093.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550.
- He YF, Li BZ, Li Z, Liu P, Wang Y, et al. (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* 333: 1303–1307.
- Hemberger M, Dean W, Reik W (2009) Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat Rev Mol Cell Biol* 10: 526–537.
- Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452: 215–219.
- Widschwendter M, Berger J, Hermann M, Muller HM, Amberger A, et al. (2000) Methylation and silencing of the retinoic acid receptor-beta2 gene in breast cancer. *J Natl Cancer Inst* 92: 826–832.
- Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118–127.