BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 11 2012, pages 1487–1494
doi:10.1093/bioinformatics/bts170

Gene expression

Advance Access publication April 6, 2012

Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions

Andrew E. Teschendorff[1,*] and Martin Widschwendter[2]

[1]Statistical Genomics Group, Paul O'Gorman Building, UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT and [2]Department of Womens Cancer, UCL Elizabeth Garrett Anderson Institute for Womens Health, University College London, Room 340, 74 Huntley Street, London, WC1E 6AU, UK

Associate Editor: Janet Kelso

ABSTRACT

Motivation: The standard paradigm in omic disciplines has been to identify biologically relevant biomarkers using statistics that reflect differences in mean levels of a molecular quantity such as mRNA expression or DNA methylation. Recently, however, it has been proposed that differential epigenetic variability may mark genes that contribute to the risk of complex genetic diseases like cancer and that identification of risk and early detection markers may therefore benefit from statistics based on differential variability.

Results: Using four genome-wide DNA methylation datasets totalling 311 epithelial samples and encompassing all stages of cervical carcinogenesis, we here formally demonstrate that differential variability, as a criterion for selecting DNA methylation features, can identify cancer risk markers more reliably than statistics based on differences in mean methylation. We show that differential variability selects features with heterogeneous outlier methylation profiles and that these play a key role in the early stages of carcinogenesis. Moreover, differentially variable features identified in precursor non-invasive lesions exhibit significantly increased enrichment for developmental genes compared with differentially methylated sites. Conversely, differential variability does not add predictive value in cancer studies profiling invasive tumours or whole-blood tissue. Finally, we incorporate the differential variability feature selection step into a novel adaptive index prediction algorithm called EVORA (epigenetic variable outliers for risk prediction analysis), and demonstrate that EVORA compares favourably to powerful prediction algorithms based on differential methylation statistics.

Conclusions: Statistics based on differential variability improve the detection of cancer risk markers in the context of DNA methylation studies profiling epithelial preinvasive neoplasias. We present a novel algorithm (EVORA) which could be used for prediction and diagnosis of precursor epithelial cancer lesions.

Availability: R-scripts implementing EVORA are available from CRAN (*www.r-project.org*).

Contact: a.teschendorff@ucl.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 13, 2012; revised on March 8, 2012; accepted on April 2, 2012

1 INTRODUCTION

Statistical testing in the context of large omic datasets (e.g. mRNA expression or DNA methylation arrays) has relied on the assumption that the most relevant features are those which differ most significantly in terms of their mean levels (Bengtsson *et al.*, 2001). Thus, parametric testing using *t*-tests or regularized versions thereof has been the standard approach for identifying and ranking features of interest (Smyth, 2004; Tibshirani *et al.*, 2002; Tusher *et al.*, 2001; Wettenhall and Smyth, 2004). This paradigm of testing for differences in the means of two distributions has been enormously successful, as demonstrated for instance by the popularity of algorithms and R-packages such as *limma* or *pamr* (Tibshirani *et al.*, 2002; Wettenhall and Smyth, 2004). However, recent insights obtained in cancer epigenomics suggest that identifying features that differ in terms of the second moment of a distribution, i.e. variability, may also be as relevant or important for understanding or predicting disease phenotypes (Feinberg and Irizarry, 2010; Feinberg *et al.*, 2010; Issa, 2011; Jaffe *et al.*, 2012). Indeed, it has been proposed that certain genes which are prone to stochastic epigenetic variation, may contribute to the risk of complex genetic diseases like cancer and that exposure to environmental risk factors may underlie much of this stochastic variation (Feinberg and Irizarry, 2010; Feinberg *et al.*, 2010). Supporting this view, a recent study has shown that features that are differentially methylated between cancer and normal tissue are generally also much more variable in cancer tissue itself, suggesting that the relevant features could be identified using the concept of differential variability (Hansen *et al.*, 2011).

In this manuscript we formally demonstrate, in the context of DNA methylation studies profiling precancerous conditions, that statistics based on differential variability identifies true positives more reliably than statistics based on differential methylation. We show how this can be attributed to the importance of heterogeneous outlier methylation profiles where the outliers define a small proportion of samples in the precancerous phenotype. Conversely, we also demonstrate that differential variability underperforms whenever outlier methylation profiles are conspicuously absent, as is the case for invasive cancer and whole-blood (WB) tissue. We show that the outlier features present in the early precursor cancer lesions are more heavily enriched for developmental genes than differentially methylated ones, underpinning their biological importance. Finally, we incorporate the differential

[*]To whom correspondence should be addressed.

variability feature selection step into an adaptive index classification algorithm (Tian and Tibshirani, 2011), thus building a novel prediction algorithm and demonstrating its added value over two powerful classification algorithms that select features based only on differential methylation statistics. We illustrate all of these results mainly in the context of cervical cancer (CC), using four DNA methylation datasets (a total of 311 samples, all profiled at 27578 CpG sites) encompassing all stages of cervical carcinogenesis.

2 METHODS

2.1 Infinium 27k methylation beadchips

All samples used in this article were profiled using Illumina's Infinium Human Methylation 27k Beadchips (Bibikova *et al.*, 2009). The Beadchips interrogate the methylation status of 27578 CpGs. In all cases the data were subject to a quality control and normalization procedure as described in our previous work (Teschendorff *et al.*, 2010). This procedure removes poor quality probes and assesses the degree of technical variation compared with biological variability using a framework based on singular value decompositions (Teschendorff *et al.*, 2009, 2010). Let i denote the CpG and s the sample. The normalized methylation values of the CpGs follow an approximate β-valued distribution, with β constrained to lie between 0 (unmethylated locus) and 1 (methylated). This follows from the definition of β as the ratio of methylated to combined intensity values i.e.

$$\beta_{is} = \frac{M_{is}}{U_{is}+M_{is}+e} \tag{1}$$

where U_{is} and M_{is} are the unmethylated and methylated intensity values of the CpG (averaged over bead replicate probes) in sample s, and e is a small correction term to regularize probes of low total signal intensity (i.e. probes with $U_{is}+M_{is}\approx 0$ after background subtraction).

2.2 DNA methylation datasets

The four main datasets considered in this work (DataSets1–4) are independent datasets. Thus, although the samples represent three different stages in carcinogenesis, they are all from different women.

DataSet 1: LBC1 (48 samples) A total of 48 liquid-based cytology (LBC) samples (all HPV+): 24 of normal histology, and 24 were CIN2+ (cervical intraepithelial neoplasia of grade 2 or higher). Normal and CIN2+ samples were age-matched. This dataset is available from Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo/) accession number GSE37020 (to be determined).

DataSet 2: LBC2 (48 samples) A total of 30 LBC samples (HPV negatives and positives) with normal cytology and 18 LBC CIN2+ samples (all HPV positive), as described in (Teschendorff *et al.*, 2010). Normals and CIN2+ samples were age-matched. This dataset is available from GEO (www.ncbi.nlm.nih.gov/geo/) accession number GSE20080.

DataSet 3: CC (63 samples) A total of 63 cervical tissue samples: 48 CCs, 15 normals. Normals and cancers were age-matched. This dataset is available from GEO (www.ncbi.nlm.nih.gov/geo/) accession number GSE30760.

DataSet 4: ARTISTIC (ART) cohort (152 samples) All samples represent cytologically normal cells from the uterine cervix of 152 women collected in a nested case-control study within ARTISTIC [a Randomized Trial of HPV Testing in Primary Cervical Screening (Kitchener *et al.*, 2009a, b)]. Within the nested case control study, 75 women developed a cervical intraepithelial neoplasia of grade 2 or higher (CIN2+) within a three year follow-up period (cases), whereas 77 women did not develop any abnormal cytology (controls). Cases and controls were matched for age and HPV status [92 were HPV positive (44 cases/48 controls) and 60 were HPV negative (31 cases/29 controls)]. This dataset is available from GEO (www.ncbi.nlm.nih.gov/geo/) accession number GSE30760.

DataSets 5–7 In addition to the previous four sets, we also include two additional DNA methylation studies encompassing relatively larger numbers of normal and cancer epithelial tissue as well as one dataset consisting of WB samples: DataSet 5: 23 normal breast and 113 breast cancer specimens (Zhuang *et al.*, 2012), DataSet 6: 23 normal endometrial and 64 endometrial cancer samples (Zhuang *et al.*, 2012), and DataSet 7: 148 WB samples from healthy women and 113 WB samples from women with ovarian cancer (Teschendorff *et al.*, 2009).

2.3 Feature selection using differential variability

Given a feature (i.e. CpG) with DNA methylation profile $\vec{\beta}\in(0,1)^n$ where n is the number of samples, differential variability tests for differences in variability between two phenotypes. Thus, if $\vec{\beta}_1$ and $\vec{\beta}_2$ refer to the methylation profiles over two phenotypes '1' and '2', we ask if the respective variances of $\vec{\beta}_1$ and $\vec{\beta}_2$, are statistically different or not. The standard statistical test for testing for differential variability is known as Bartlett's test, or its non-parametric equivalent, the Levene test (Snedecor and Cochran, 1967). In principle, the Levene test should be preferable since it does not assume that the data are normally distributed and indeed it is more robust to outliers. However, in the present context, outliers are the biologically interesting features (see later) and we seek a test that can identify differential variability *due to outliers*. Hence, we here use Bartlett's test as a means of selecting differentially variable features where the differential variability is driven by a potentially small number of outliers. Further justification for this choice is provided *a posteriori* by demonstrating that this indeed leads to improved feature selection.

Let s_1^2 and s_2^2 denote the variances of phenotypes 1 and 2, and let (n_1,n_2) denote the number of samples in each phenotype (i.e. $n=n_1+n_2$). Then, under the null hypothesis ($s_1^2=s_2^2$), the Bartlett test statistic, B,

$$B=\frac{(n-2)\log s^2-\sum_{i=1}^{2}(n_i-1)\log(s_i^2)}{1+\frac{1}{3}\sum_{i=1}^{2}\frac{1}{n_i-1}-\frac{1}{n-2}} \tag{2}$$

is χ^2-distributed with 1 degree of freedom, and where $s^2=1/n-2\sum_{i=1}^{2}(n_i-1)s_i^2$. It is from this statistic that a $\underline{P}$-value is derived. For convenience and interpretation, we also consider the b-statistic, defined as the log ratio of the variances, i.e. $b=\log\{s_2^2/s_1^2\}$.

2.4 Epigenetic variable outliers for risk prediction analysis

Epigenetic variable outliers for risk prediction analysis (EVORA) is similar to an adaptive index prediction algorithm (Tian and Tibshirani, 2011) and is based on the following three biologically motivated hypotheses:

- Relevant DNA methylation features (i.e. CpGs capable of predicting disease phenotypes) may be identified more accurately by comparing the variance in methylation between phenotypes, as proposed in (Feinberg and Irizarry, 2010; Hansen *et al.*, 2011). We refer to these CpGs generally as true positives, 'risk CpGs' or 'diagnostic CpGs' depending on biologically context.
- Risk CpGs exhibit outlier methylation profiles and can be identified more accurately using differential variability. This is motivated by the hypothesis that much of the epigenetic variation is stochastic (Feinberg and Irizarry, 2010).
- The cancer risk score of an individual sample is proportional to the number of risk CpGs that constitute significant methylation outliers (specifically hypermethylation) in that given sample. We call these outliers, 'methylation hits'. This proportionality assumption is reasonable as it reflects the degree of deviation from normal baseline methylation levels in the healthy phenotype.

To translate these model assumptions into a prediction tool, we need to be able to (i) identify risk CpGs and samples that constitute outliers for these candidate risk CpGs; and (ii) a statistical method for assigning risk to each sample, and preferably one which is robust and independent of the scale used. Scale independence is important to guarantee that classification thresholds are generalizable to independent datasets. To address problem (i) we propose to transform the DNAm data matrix so that outliers can be identified in an objective manner independently of the scale used. To achieve this, we use the COPA (Cancer Outlier Profile Analysis) transformation, which was first used in the context of gene expression data to identify candidate gene fusions with outlier gene expression (Tomlins *et al.*, 2005). Specifically, for each CpG with methylation profile $\vec{\beta}$ we transform it to a COPA-profile

$$\beta_s^C = \frac{\beta_s - \text{median}(\vec{\beta})}{\text{mad}(\vec{\beta})} \tag{3}$$

where s denotes the sample, and where *mad* denotes the median absolute deviation. In other words, we subtract out the median of a profile, then find the median of the non-zero absolute deviations, and finally divide the median centred profile by this number. Although this transformation is linear, the use of the median and *mad* makes the transformation robust to outliers. Moreover, the COPA-transformed values are independent of the scale used and for a given threshold one may now define samples that constitute outliers for that specific CpG profile. Note again that because most of the CpGs on the 27k arrays map to promoters and that the great majority of these are unmethylated in normal healthy tissue, that the interesting methylation outliers are those which show higher methylation in the transformed phenotype. Thus, we are here interested in samples with high (positive) COPA scores. We note that in the case of the denser 450k Illumina arrays (Sandoval *et al.*, 2011) where probes are not biased to promoter regions and for which therefore hypomethylation may play an equally important role, large negative COPA scores would also be of interest. Thus, EVORA could be generalized to this scenario by taking the absolute value of the COPA scores. To address problem (ii) we propose the EVORA algorithm, which consists of the following steps. We assume that samples can be labelled by a binary phenotype with 'N' labelling the normal state and 'T' indicating the transformed phenotype (e.g. a prospective CIN2+, a CIN2+ or CC):

(1) Perform the COPA transformation on the whole DNAm data matrix. Phenotype information is obviously not used in this step.

(2) Start an internal 10-fold cross-validation. Each fold consists of splitting the samples up into a training and test set, ensuring that training and test sets contain equal relative proportions of the two phenotypes, and that each sample is used only once across all test sets.

(3) Using the original β-valued data matrix, identify and rank candidate risk CpGs using Bartletts test in the training set (i.e. identify CpGs more variable in phenotype 'T', we call these CpGs 'hyperV DVCs'). Perform FDR (false discovery rate) analysis to determine if there is a sufficient number of CpGs passing an appropriate threshold. Typically we require on the order of hundreds of CpGs to pass an FDR < 0.05 threshold, however, the threshold may be relaxed to FDR < 0.3 to ensure this number.

(4) For each candidate risk CpG and for a range of choices of COPA thresholds, transform the COPA methylation profile to a binarized EVORA profile, in which each sample is scored as 1 (if the COPA value for that sample is larger than the current threshold) or 0 (if the COPA value is less than the threshold).

(5) For each sample in the test set calculate the fraction of risk CpGs which are outliers (i.e. number of 1s). This fraction is the risk index of that sample and is dependent on the COPA threshold and on the number of top ranked risk CpGs included.

(6) Repeat Steps 3–5 for each fold. Thus, each sample acts as a test sample once and is assigned a risk score.

(7) Compute the area under the ROC curve (AUC) of the resulting risk scores at each COPA threshold and for different numbers of top ranked risk CpGs (typically starting at 50 and ending at 1000 or the maximum number obtained from Step 3).

(8) Find the COPA threshold and number of risk CpGs that optimizes the AUC over the internal cross-validation.

(9) Having identified the optimal parameters, the risk score of an independent sample is obtained as the fraction of selected risk CpGs that have this sample as an outlier according to the optimal COPA threshold.

We make several notes about this procedure:

(i) Having identified the optimal number of risk CpGs, n^*, to be included, the exact composition of this final list of risk CpGs may be constructed by considering the union of all sets of risk CpGs obtained in the internal cross-validation. Specifically, for the union set of such risk CpGs we count how often a given CpG is present in each run and rank CpGs according to how often they are chosen. We declare the final risk CpG set as the top n^* of this ranked list.

(ii) Since the risk score is just the percentage of methylation hits in a given sample, different samples may be assigned the same risk score. To discriminate samples with tied risk scores we computed a mean methylation score over the risk CpGs using the original β-values. The mean methylation value of each of these tied samples was then renormalized to lie in the open interval (0,1) using the transformation $\beta \rightarrow \beta - \beta_{\text{min}}/\beta_{\text{max}} - \beta_{\text{min}} \mp \epsilon$, where ϵ is a small offset term to ensure that the scores lie in the open interval (i.e. values 1 and 0 are to be excluded since otherwise this would induce ties with samples which have neighbouring risk scores). This whole procedure therefore ranks samples first according to percentage of methylation hits and resolves ties using a mean methylation score.

(iii) It is important to stress that the EVORA risk score is an outlier risk score, i.e. it is derived from a scale (the COPA-scale) that allows improved identification of outliers. This also means that, potentially, the optimal COPA threshold is trained on a specific type of outlier profile. Mathematically, outlier profiles are bi-modal (or multi-modal) and these can be of two types: (i) profiles where the outlier group is made up of a relatively small number of samples, which we call heterogenous outliers; and (ii) more homogeneous profiles where the outlier group has substantially more samples and is more similar to the size of the normal group (i.e. the 'N' phenotype). Thus, if the training and test datasets differ significantly in terms of the type of outlier profiles they exhibit, this could compromise the performance of EVORA. In such a scenario, the COPA threshold can be transformed back into the β-value scale, and the score in the independent sample computed using this threshold on the original β-values.

3 RESULTS

Based on the insights of previous epigenetic studies (Feinberg and Irizarry, 2010; Feinberg *et al.*, 2010), we hypothesized that differential variability in DNA methylation would play a particularly important role in the context of early carcinogenesis. Thus, to formally demonstrate that differential variability identifies more biologically relevant DNA methylation features than statistics based on differential methylation, we devised an objective training/evaluation set strategy in the context of CC, a cancer for which the cell of origin is known and for which access to early precursor lesions is possible through routine large scale screening programs (Kitchener *et al.*, 2009a, b). Specifically, we analyzed DNA methylation profiles for a total of 311 samples, including normal tissue, normal cells preceding a high-grade

Table 1. Study design: the breakup of samples in each dataset (ART, LBC1, LBC2 and CC) according to cytology/histology (N = normal, preCIN2+ = precursor CIN2+, CIN2+ = cervical intraepithelial neoplasia of grade 2 or higher, CC = cervical cancer) is shown

	ART[a]	LBC1[a]	LBC2	CC
N	77	24	30	15
preCIN2+	75	0	0	0
CIN2+	0	24	18	0
Cancer	0	0	0	48

[a] These studies were used for discovery/training since they were balanced in terms of phenotypes. We note that the 311 samples are all independent (i.e. from 311 different women).

cervical intraepithelial neopasia (CIN2+) (i.e. precursor CIN2+ lesions), CIN2+ samples and (invasive) CC, encompassing three progressive stages and drawn from four independent studies (Section 2, Table 1). Thus, our strategy was to pick as discovery/training sets those cohorts representing the two earliest stages of CC (precursor CIN2+ and CIN2+ lesions both of which are non-invasive), and to use cohorts representing similar or more advanced stages as evaluation/test sets. In one context we used LBC1, consisting of a balanced number of normal (24) and CIN2+ (24) samples, as the discovery/training set, and used sets LBC2 and CC for evaluation/testing. In the second context, we used the ARTISTIC (ART) cohort representing a balanced set of normal (77) and precursor CIN2+ (75) samples for discovery/training, and used LBC1, LBC2 and CC as evaluation/test sets.

3.1 Differential variability identifies differentially methylated heterogeneous outlier features

Our first task is to demonstrate that differential variability selects a distinct but still relevant set of features, compared with ordinary statistics (e.g. t-tests) which rank features according to the degree of differential methylation. Accordingly, we took LBC1 as our discovery set to identify features (CpGs) of interest. In one case we used t-test P-values to rank all CpGs that passed quality control (over 23 000 of them) according to how well they could discriminate normal from CIN2+ status differentially methylated CpGs-DMCs). In the second case, we ranked CpGs according to differential variability between the same two phenotypes using Bartlett's test (differentially variable CpGs-DVCs) (Section 2). In both cases, histograms of P-values and associated FDR estimation using q-values (Storey and Tibshirani, 2003) indicated a substantial number of CpGs that were associated with CIN2+ status (Supplementary Figure S1). Among the top 500 DMCs there was a highly significant skew (77%) towards hypermethylation in the CIN2+ phenotype. Similarly, among the top 500 DVCs, the great majority (87%) were hypervariable in the CIN2+ phenotype. Only 2 CpGs overlapped between these two lists of 500 CpGs, a significant underenrichment [Fisher-test odds ratio (OR) = 0.18, $P < 0.05$], indicating that testing for differential variability is indeed selecting different features to those derived from ordinary t-tests. This, however, does not mean that DVCs are not differentially methylated. Indeed, a scatterplot of t-test P-values of the top 500 DVCs clearly showed that many are indeed differentially methylated, albeit obviously not as highly ranked as their t-test derived counterparts (Fig. 1A). Thus, differential variability is selecting a different subset of differentially methylated features. Importantly, almost all hypervariable DVCs had positive t-statistics, meaning that they are more methylated in CIN2+ samples (Fig. 1B). In contrast, DMCs were on the whole less variable in the CIN2+ phenotype (Fig. 1B). The different feature selection obtained under Bartlett's test as compared with t-tests can be explained by marked differences in the shapes of the selected methylation profiles. Indeed, DVCs generally exhibited an outlier structure with only a subset of the CIN2+ samples showing increased methylation (typical increments of at least 20%) (Fig. 1C). Thus, Bartlett's test provides a useful algorithm for identifying CpGs that exhibit a particular type of outlier profile. In contrast, and as expected, t-tests were observed to select for CpGs that showed more homogeneous differential methylation changes (Fig. 1D).

3.2 Differential variability in precursor and preneoplastic lesions improves the predictive value of CIN2+ and cancer

Next, we asked which set of ranked CpGs (i.e. DVCs or DMCs) identified more true positives, defined as features whose methylation level correlates with CIN2+ status in an independent test set. Given the significant skew towards hypermethylation and hypervariability in the CIN2+ phenotype, we focused our analysis on the top ranked 500 hypervariable DVCs and top 500 hypermethylated DMCs from set LBC1. We note that the skew towards hypervariability and hypermethylation in the transformed phenotype was observed in almost all datasets, justifying our focus on hypervariability and hypermethylation (Supplementary Table S1). As evaluation set we used the independent cohort LBC2, which also consists of normal and CIN2+ samples (Section 2, Table 1). To estimate the strength of correlation to CIN2+ status in this test set we used the t-statistic. We observed that the test set t-statistics of the top 500 DVCs identified in the training set (LBC1) was higher than the corresponding test set t-statistics of the top 500 DMCs (Fig. 2A). Given that DMCs were selected using t-tests, it is therefore remarkable that in the evaluation set, DVCs outperformed DMCs in terms of the t-statistic. Furthermore, declaring a true positive feature as one with a nominal t-test P-value < 0.05 in the evaluation set, the positive predictive value (PPV) of DVCs was almost twice as large as those of DMCs (Table 2). Using a more stringent Benjamini–Hochberg corrected P-value threshold of 0.05, results were robust for 4 of the 5 comparisons (Supplementary Table S2). As expected, DVCs also showed higher b-statistics (Bartlett-test) in the evaluation set than DMCs (Supplementary Figure S2). All these results, therefore, indicate that differential variability identifies true positives more reliably than differential methylation, at least in the context of preinvasive cancer lesions.

To further strengthen the case for differential variability, we considered a third dataset (CC, Section 2), consisting of 48 CC tissues and 13 normal cervix specimens. As before, we computed for each of the top 500 hypervariable DVCs and top 500 hypermethylated DMCs from dataset LBC1, the t-statistic in this third set. Remarkably, once again DVCs exhibited higher t-statistics than DMCs (Fig. 2A) and their PPV was also much higher (Table

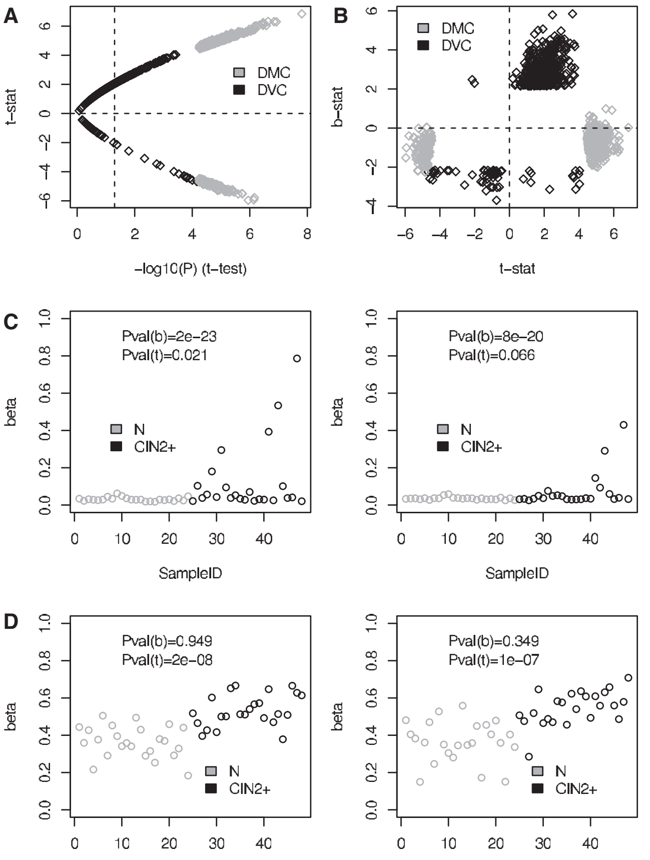

Fig. 1. Differential variability selects for differentially methylated outlier features which are distinct from those obtained using *t*-statistics. (**A**) Plots of *t*-statistics (*y*-axis) versus −log10(*P*-values) (*x*-axis) for the top 500 DMCs and top 500 DVCs in dataset LBC1. Vertical dashed line indicates the nominal threshold *P*=0.05. The 500-th top ranked DMC had a FDR of 0.003. (**B**) Scatterplot of Bartlett-test b-statistics (*y*-axis) against *t*-statistics (*x*-axis) of the top 500 DMCs and top 500 DVCs in dataset LBC1. The b-statistic is the logarithm of the ratio of the variances in CIN2+ to the variance in the normal phenotype. (**C**) Typical methylation profiles of top ranked DVCs cg02440177 (left) and cg08876932 (right). (**D**) Typical methylation profiles of top ranked DMCs cg06948937 (left) and cg15537850 (right). In (C) and (D), *y*-axis labels β-value, *x*-axis the sample. Phenotypes (*N* = normal,CIN2+ = cervical intraepithelial neoplasia of grade 2 or higher) are as indicated. *P*-values obtained from the Bartlett test statistic (b) and *t*-test statiistic (*t*) are given

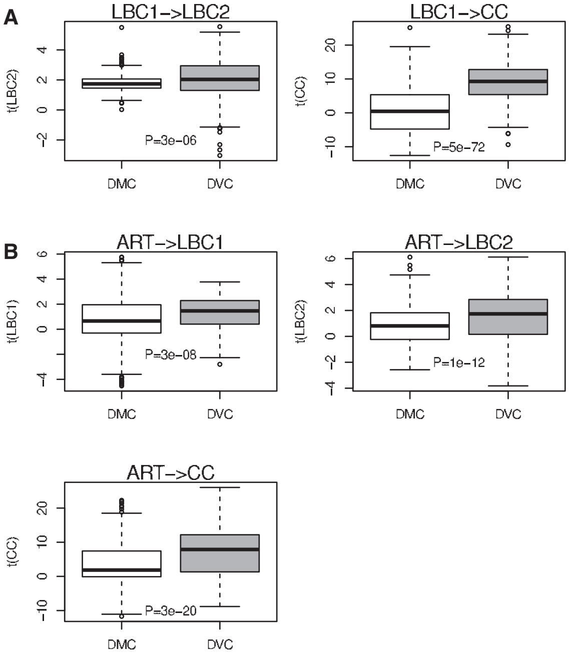

Fig. 2. DVCs selected in a training set achieve higher *t*-statistics in test sets compared with DMCs. (**A**) *t*-statistics of the top 500 hypermethylated DMCs and top 500 hypervariable DVCs identified from the training set LBC1, in the independent datasets LBC2 (30 normals and 18 CIN2+) and CC (15 normals and 48 CCs). (**B**) *t*-statistics of the top 500 hypermethylated DMCs and top 500 hypervariable DVCs identified from the training set ART, in the independent datasets LBC1 (24 normals and 24 CIN2+), LBC2 (30 normals and 18 CIN2+) and CC (15 normals and 48 CCs). In all cases, *P*-values are from a Wilcoxon rank sum test comparing the *t*-statistics

Table 2. The PPV of the top ranked 500 DMCs and top 500 DVCs derived from the discovery set in the evaluation/test sets, as indicated

Discovery	LBC1	LBC1	ART	ART	ART
test set	LBC2	CC	LBC1	LBC2	CC
DMCs	0.29	0.42	0.23	0.19	0.48
DVCs	**0.48**	**0.90**	**0.33**	**0.44**	**0.73**

In the case where feature selection was done in LBC1, LBC2 and CC were used as evaluation sets. In the case where feature selection was done in the ART set, evaluation was done in LBC1, LBC2 and CC.

2). Given that CIN2+ precedes CC, this result clearly indicates that differential hypervariable CpGs in CIN2+ lesions are more likely to be hypermethylated in CC than CpGs identified on the basis of differential methylation. DVCs identified in LBC1 were also much more variable in CC compared with normal tissue than DMCs (Supplementary Figure S2).

As another example, we considered a fourth dataset (ART, Section 2) consisting of 152 prospectively collected cervical smear samples, all of normal cytology at the time of sample draw. Out of these 152 samples, 75 developed a CIN2+ within a 3 year follow-up period. Since these samples represent the earliest stage of cervical carcinogenesis, we used this dataset as our new training set to identify DVCs and DMCs associated with prospective CIN2+ status and to then compare their *t*-statistics in the three independent sets (LBC1, LBC2 and CC). First, we only observed a very weak genome-wide statistical significance when using *t*-tests, in contrast to DVCs which did show genome-wide significance levels (Supplementary Figure S3). Next, we selected the top 500 hypermethylated DMCs and top 500 hypervariable DVCs from the training set. Similar to the results obtained previously, DVCs exhibited significantly higher *t*-statistics and b-statistics in all three evaluation sets (Fig. 2B, Supplementary Figure S2). In line with this, PPV values in the evaluation sets were also higher for DVCs than DMCs (Table 2).

Table 3. Prediction performance in the test sets as measured by the Area Under the Curve (AUC), plus 95% CIs

Training test set	LBC1 LBC2	LBC1 CC	ART LBC1	ART LBC2	ART CC
PAMR	0.62	0.75	**0.69**	0.64	0.83
	(0.46–0.78)	(0.63–0.86)	(0.53–0.84)	(0.49–0.80)	(0.73–0.93)
SPCA	0.61	0.94	0.66	0.73	**0.94**
	(0.44–0.78)	(0.88–0.99)	(0.51–0.81)	(0.59–0.87)	(0.88–1.00)
EVORA	**0.87**	**1**	**0.69**	**0.87**	**0.94**
	(0.77–0.97)	NA	(0.54–0.85)	(0.67–0.92)	(0.89–0.99)

EVORA is compared with a popular nearest shrunken centroid classification algorithm (PAMR) and SPCA with a maximum of three components. In both PAMR and SPCA, feature selection is done via differential methylation statistics. In the case where training was done in LBC1, LBC2 and CC were used as test sets. In the case where training was done in the ART set, testing was done in LBC1, LBC2 and CC.

Table 4. Comparison of PCGT gene enrichment OR, 95% CIs and *P*-values of enrichment (*P*) among the top 500 hypervariable DVCs and the top 500 hypermethylated DMCs, in each of the four datasets (ART, LBC1, LBC2 and CC)

DataSet	hyperV DVCs OR (95% CI)	*P*	hyperM DMCs OR (95% CI)	*P*
ART	6.2 (4.8–8.1)	2e-33	3.1 (2.2–4.1)	6e-11
LBC1	9.3 (7.2–11.8)	1e-54	1.5 (0.9–2.2)	0.08
LBC2	7.9 (6.1–10.2)	8e-43	5.4 (4.1–7.1)	2e-25
CC	6.8 (5.2–8.8)	2e-36	6.7 (5.1–8.7)	3e-33

CI, confidence interval.

3.3 Improved prediction of CIN2+ and cancer status using EVORA

To further demonstrate the importance of differential variability, we next focused on the task of prediction. We incorporated the differential variability feature selection step into a novel classification framework based on an adaptive index algorithm (Tian and Tibshirani, 2011). We call the resulting prediction algorithm, EVORA (Section 2). Briefly, EVORA selects features (‘risk CpGs’) using Bartlett’s test for differential variability, and subsequently assigns a ‘risk’ score to an independent sample by counting the fraction of risk CpGs which constitute methylation outliers in that sample. A CpG constitutes a methylation outlier (or ‘hit’) in that sample if the CpG exhibits abnormally high methylation compared with the median value across all samples as determined by the COPA transformation (Tomlins *et al.*, 2005) and a threshold parameter (Section 2). The optimal threshold and number of risk CpGs to be included are determined by an internal cross-validation (Section 2).

We trained EVORA on the LBC1 and ART cohorts, and subsequently tested the predictions in the independent cohorts representing similar or more advanced stages (Table 1). Predicted EVORA risk scores in the evaluation sets were correlated to case/control status using the AUC. To benchmark EVORA, we compared it with a popular nearest shrunken centroid prediction algorithm (PAMR) (Tibshirani *et al.*, 2002) and SPCA (Bair and Tibshirani, 2004), which are both based on feature selection using differentially methylation statistics. In line with our PPV analysis, we observed that EVORA obtained higher AUC values than PAMR and SPCA in 3 of the 5 comparisons, and in none of the 5 did it underperform (Table 3). In several instances the differences in AUC over PAMR or SPCA were substantial (Table 3). Heatmaps over the DVCs selected by EVORA confirmed the increased variability in the transformed phenotypes and that EVORA risk scores correlate with CIN2+ and cancer status (Supplementary Figure S4).

3.4 Enrichment of PRC2 target genes is stronger among DVCs than DMCs

Next, we asked if the different feature selection due to DVCs is reflected by differences in gene ontology enrichment. Since it is known that developmental genes and in particular genes marked by the PRC2 repressive complex [polycomb group targets (PCGTs)] are preferentially hypermethylated in cancer relative to normal tissue (Lee *et al.*, 2006; Ohm *et al.*, 2007; Schlesinger *et al.*, 2007; Widschwendter *et al.*, 2007), one would expect that the same class of genes would be equally enriched among DVCs. Indeed, previous studies have provided evidence that developmental genes and PCGTs are targets of increased epigenetic variability in cancer (Feinberg and Irizarry, 2010; Hansen *et al.*, 2011). To test this further in the preinvasive cancer context, we compared the enrichment OR of PCGTs among the top 500 hypervariable DVCs and top 500 hypermethylated DMCs in each of the three studies representing the earliest stages of cervical carcinogenesis (ART, LBC1 and LBC2). In all of them, we observed significantly higher PCGT enrichment OR for DVCs than DMCs (Table 4). In contrast, the PCGT enrichment OR was similar for DVCs and DMCs derived in the CC study (CC, Table 4). Thus, when differences in methylation are fairly large and more homogeneous, as is the case when comparing cancer to normal tissue (CC set) (Supplementary Figure S4), statistics based on differential variability or differential methylation lead to a similar PCGT enrichment.

3.5 Differential variability in invasive cancer or WB tissue does not improve the predictive value of cancer diagnostic markers

The improved PPV (and AUC) of DVCs derived from precursor/preinvasive cervical cancer tissue relies on the fact that in these early stages of carcinogenesis there are many outlier methylation profiles which turn out to be of biological relevance (Fig. 1A, C). We thus posited that differential variability would be less useful in contexts where no such methylation outlier profiles are present. Specifically, this would be the case for (invasive) cancer, because cancer is characterized by a more homogenous methylation of CpG sites within promoters and CpG islands (see Supplementary Figure S4). Similarly, cancer diagnostic markers derived from WB samples do not exhibit outlier methylation profiles, as shown by us previously (Teschendorff *et al.*, 2009) (Supplementary Figure S5).

To test our hypothesis, we collected three additional datasets (DataSets 5–7, Section 2) encompassing relatively larger numbers of normal and cancer tissue (Zhuang *et al.*, 2012), and WB tissue from (ovarian) cancer patients and healthy controls (Teschendorff *et al.*, 2009). The larger number of normal samples in these three datasets allowed a training-test set partition strategy to be used to compare the cross-validation statistics based on selecting either DMCs or DVCs. In line with our expectations, the statistics associated with DVCs were significantly lower than those of DMCs in the WB study (Supplementary Figure S5) as well as in the two invasive cancer studies (Supplementary Figure S6).

4 DISCUSSION AND CONCLUSIONS

Although several studies have proposed that stochastic epigenetic variation is an intrinsic characteristic of the cancer phenotype (Feinberg and Irizarry, 2010; Feinberg *et al.*, 2010), so far only one study has investigated the role of differential variability in identifying cancer-associated DNA methylation markers (Hansen *et al.*, 2011). Indeed, in (Hansen *et al.*, 2011) it was shown that regions which are differentially methylated between cancer and normal tissue constitute regions which are more prone to variability in the cancer phenotype itself. However, no study has formally demonstrated the added value of using differential variability to identify features that may indicate the risk of a transformed phenotype.

In this work, we fill this gap by formally demonstrating that differential variability can be used to identify more reliably CpGs whose methylation levels are associated with a transformed phenotype. Specifically, we have shown that CpGs which were hypervariable in the transformed phenotype were more likely to be associated with the transformed phenotype in independent datasets, compared with CpGs which were selected based on differential methylation. We showed this in the context of two different training sets reflecting two different stages of early oncogenesis. In the case of the ARTISTIC cohort, we showed how hypervariable CpGs in precursor CIN2+ lesions were stronger indicators of CIN2+ status and CC in independent datasets. Similarly, hypervariable CpGs in CIN2+ samples were also stronger predictors of CIN2+ and CC status in independent cohorts. In this context, it is remarkable that features which were selected using t-tests had lower t-statistics in the evaluation sets than features which were selected using differential variability. This last result attests to the importance of differential variability in the early stages of carcinogenesis.

It is important to point out, however, that differential variability may not improve the identification of relevant markers in other contexts. Indeed, we have seen that the identification of cancer diagnostic markers derived from studies profiling invasive cancers (a more advanced stage of carcinogenesis) does not benefit from considering differential variability. Similarly, using differential variability to identify cancer diagnostic markers from WB tissue also yielded a reduced predictive value. We can attribute the reduced performance of DVCs in these specific contexts to the fact that methylation outliers play a less fundamental role in the more advanced stages of an epithelial cancer or, in the case of WB, that cancer diagnostic markers reflect small changes in the blood cell type composition (Teschendorff *et al.*, 2009). In these contexts, statistics based on differential methylation are entirely appropriate because the associated methylation profiles in the transformed phenotype exhibit a much more homogeneous type of bi-modality, an assumption which is implicit when using, e.g. t-statistics. In contrast, the relevant methylation profiles in the earliest stages of cancer are characterized by outlier profiles, where the bi-modality and hypervariability are driven by a relatively small number of samples in the transformed phenotype, and we have seen that in this context differential variability not only improves the sensitivity of detecting such markers, but that it also improves the predictive value. Attesting to the biological importance of DVCs, we also observed that these were significantly more enriched for PCGTs than DMCs, and we note that this result too was restricted to the two earliest stages of cervical carcinogenesis.

In this manuscript, we also incorporated the concept of differential variability into a novel prediction algorithm called EVORA. EVORA uses differential variability to select features, and then uses an adaptive index algorithm over these features to assign a risk score to independent samples. Comparison of EVORA to two popular and powerful classification algorithms that are based on differential methylation statistics (PAMR, SPCA), showed that EVORA performed favourably in predicting CIN2+ and cervival cancer, further confirming the higher predictive power of DVCs over DMCs. Once again, this improved performance is restricted to the situation where the training is done in either precursor or established CIN2+ samples, representing the earliest stages of carcinogenesis.

It is important to emphasize that feature selection using differential variability is the key reason why EVORA compares favourably to prediction algorithms like PAMR or SPCA. Indeed, the relative PPVs obtained (Table 2) are in line with the relative AUC values (Table 3), indicating that the precise construction of the risk score is of less importance. Indeed, we have verified that an alternative construction of the risk score using an average methylation over the hypermethylated DVCs leads to a very similar AUC performance (data not shown). Although there are many other alternative statistical procedures one could use to compute the risk score, we have here advocated a construction that is based on a linear scale transformation (COPA) (Tomlins *et al.*, 2005) of the CpG methylation profile. The COPA transformation is robust to outliers, thus allowing a more objective framework in which to identify these. A further compelling reason for using the COPA framework, however, is that it easily allows for hypervariability associated with either hyper or hypomethylation. In fact, whereas methylation outliers associated with hypermethylation would exhibit large positive COPA scores, outliers associated with hypomethylation would exhibit large negative scores. Thus, EVORA could be easily extended to optimize a threshold over the absolute COPA scores, with the resulting risk score effectively counting both hypermethylated as well as hypomethylated outliers. The EVORA risk score could therefore be interpreted as an 'epigenetic instability index', a notion which has proved useful in the context of predicting clinical outcome in cancer DNA methylation studies (Zhuang *et al.*, 2012). Nevertheless, in this manuscript we have focused on features which become hypermethylated and hypervariable in the transformed phenotype. This is justified by our observation that in the discovery/training sets there were substantial skews towards hypermethylation and hypervariability. This skew may reflect the choice of the Infinium 27k methylation platform where most CpGs are located within promoters, which are usually unmethylated in the normal physiological state. It will therefore be interesting to explore if differential variability can also improve the detection of biologically relevant DNA methylation markers in the context of hypomethylation. To address this, however, will require denser methylation arrays [e.g. the 450k Infinium methylation platform (Sandoval *et al.*, 2011)] where probes are much less biased to promoter regions. Thus, although the EVORA algorithm implemented here is mainly aimed at 27k studies, or studies that wish to restrict to promoter regions, EVORA could be easily generalized as explained above to include differential variability associated with hypomethylation, and therefore we envisage that EVORA will also be useful in the context of 450k arrays.

Although the insights presented here have been obtained in the context of mainly one cancer (CC), it is very likely that they will generalize to other invasive neoplasias and perhaps also to other complex genetic diseases. In this regard, it is worth pointing out that CC is the only epithelial human cancer where the normal cell of origin is easily accessible in large numbers as part of routine screening programs. Since differential variability only appears to play an important role in the early stages of carcinogenesis, it would therefore be challenging to demonstrate its added value in the context of other cancers, since for most other cancers the cell of origin or precursor cancer cells are either unknown or not easily accessible in sufficiently large numbers.

In summary, we have shown that differential variability can be used to significantly improve the sensitivity and detection of cancer risk in DNA methylation studies. Statistics and algorithms based on this novel paradigm of differential variability will undoubtedly play a major role in large-scale epigenome wide studies profiling epithelial preneoplastic tissues.

Funding: A.E.T. was supported by a Heller Research Fellowship. This work was funded by the Eve Appeal (http://www.eveappeal.org.uk/) and a grant from the UCLH/UCL Comprehensive Biomedical Research Centre project, and by the NIHR Health Technology Assessment Programme and 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.

Conflict of Interest: none declared.

REFERENCES

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, E108.

Bengtsson,H. *et al.* (2001) Identifying differentially expressed genes in cdna microarray experiments authors. *Sci. Aging Knowl. Environ.*, **2001**, vp8.

Bibikova,M. *et al.* (2009) Genome-wide DNA methylation profiling using Infinium assay. *Epigenomics*, **1**, 177–200.

Feinberg,A.P. and Irizarry,R.A. (2010) Evolution in health and medicine sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl Acad. Sci. USA*, **107**, 1757–1764.

Feinberg,A.P. *et al.* (2010) Personalized epigenomic signatures that are stable over time and covary with body mass index. *Sci. Transl. Med.*, **2**, 49ra67.

Hansen,K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.

Issa,J.P. (2011) Epigenetic variation and cellular darwinism. *Nat. Genet.*, **43**, 724–726.

Jaffe,A.E. *et al.* (2012) Significance analysis and statistical dissection of variably methylated regions. *Biostatistics*, **13**, 166–178.

Kitchener,H.C. *et al.* (2009a) Artistic: a randomised trial of human papillomavirus (hpv) testing in primary cervical screening. *Health Technol. Assess*, **13**, 1–150.

Kitchener,H.C. *et al.* (2009b) Hpv testing in combination with liquid-based cytology in primary cervical screening (artistic): a randomised controlled trial. *Lancet Oncol.*, **10**, 672–682.

Lee,T.I. *et al.* (2006) Control of developmental regulators by polycomb in human embryonic stem cells. *Cell*, **125**, 301–313.

Ohm,J.E. *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.

Sandoval,J. *et al.* (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics*, **6**, 692–702.

Schlesinger,Y. *et al.* (2007) Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet.*, **39**, 232–236.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**.

Snedecor,G.W. and Cochran,W.G. (1967) *Statistical Methods*. 6th edn. Iowa State University Press, Ames, Iowa.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Teschendorff,A.E. *et al.* (2009) An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One*, **4**, e8274.

Teschendorff,A.E. *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.

Tian,L. and Tibshirani,R. (2011) Adaptive index models for marker-based risk stratification. *Biostatistics*, **12**, 68–86.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Tomlins,S.A. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wettenhall,J.M. and Smyth,G.K. (2004) limmagui: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, **20**, 3705–3706.

Widschwendter,M. *et al.* (2007) Epigenetic stem cell signature in cancer. *Nat. Genet.*, **39**, 157–158.

Zhuang,J. *et al.* (2012) The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer. *PLoS Genet.*, **8**, e1002517.