EBioMedicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

DNA Methylation Patterns in Normal Tissue Correlate more Strongly with Breast Cancer Status than Copy-Number Variants

Yang Gao [a], Martin Widschwendter [b], Andrew E. Teschendorff [a,b,c,*]

[a] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institute for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China
[b] Department of Women's Cancer, University College London, 74 Huntley Street, London WC1E 6AU, United Kingdom
[c] UCL Cancer Institute, Paul O'Gorman Building, University College London, 72 Huntley Street, London WC1E 6BT, United Kingdom

ARTICLE INFO

Article history:
Received 16 March 2018
Received in revised form 25 April 2018
Accepted 27 April 2018
Available online xxxx

Keywords:
Cancer
Cancer-risk
Early detection
DNA methylation
Epigenetic
Copy-number
Breast cancer

ABSTRACT

Normal tissue at risk of neoplastic transformation is characterized by somatic mutations, copy-number variation and DNA methylation changes. It is unclear however, which type of alteration may be more informative of cancer risk. We analyzed genome-wide DNA methylation and copy-number calls from the same DNA assay in a cohort of healthy breast samples and age-matched normal samples collected adjacent to breast cancer. Using statistical methods to adjust for cell type heterogeneity, we show that DNA methylation changes can discriminate normal-adjacent from normal samples better than somatic copy-number variants. We validate this important finding in an independent dataset. These results suggest that DNA methylation alterations in the normal cell of origin may offer better cancer risk prediction and early detection markers than copy-number changes.

1. Introduction

Throughout life, normal cells acquire somatic alterations in the genome and epigenome, both of which are thought to contribute to the onset of neoplasia and cancer [1–11]. Mapping genetic and epigenetic changes in normal tissue at risk of neoplastic transformation is therefore critically important for understanding oncogenesis, identifying early causal drivers and for cancer risk prediction [12]. Although a number of studies have been able to link somatic mutations and copy-number-variants (CNVs) in whole blood to the future risk of hematological and solid cancers [2,4–6,13,14], analogous results for somatic alterations in the epithelial cell of origin of solid cancers have remained elusive. Indeed, identifying somatic mutations in normal tissue is technically challenging [12,15,16], with only a couple of studies having been able to associate epithelial cancer risk to somatic mutations in normal (epithelial) tissue [17,18]. In contrast, DNA methylation (DNAm) changes have been correlated to cancer risk in blood [7,19–21], are frequently observed in preneoplastic epithelial tissue [22–27], and in the context of cervical smears have allowed prospective risk prediction of a high-grade intraepithelial neoplasia independently of HPV status [25].

Two recent studies formally compared somatic mutations/CNVs to DNAm changes in their ability to predict prospective risk of gastric and esophageal cancer [17,18]. One study showed that DNAm changes may be a better risk predictor than somatic mutations, specially for gastric cancer [18], whilst the other study showed that both CNVs and DNAm changes were better than somatic mutations at predicting progression of intestinal metaplasia to gastric cancer [17]. Thus, both studies underscore the importance of DNAm changes in carcinogenesis and suggest that epigenetic alterations may be a better molecular cancer risk predictor than genetic changes. However, despite these two studies, the relative importance of genetic and epigenetic alterations for cancer risk prediction remains unclear.

Here we decided to shed further light on this outstanding question. Although comparing different types of molecular alteration as predictors of cancer risk is technically challenging due to the requirement of measuring all relevant molecular profiles in the relevant tissue and in a relatively large number of individuals, several studies have shown the feasibility of using Illumina Methylation 450 k/EPIC beadarrays to obtain high-confidence CNV calls [28–30], thus allowing at least for an objective comparison between CNV and DNAm. Here we conduct such a comparison in the context of an epithelial cancer using a cohort of 50 normal healthy breast samples, 42 age-matched normal samples collected adjacent to breast cancer, and a total of 305 invasive breast

* Corresponding author at: CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institute for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China.
E-mail address: a.teschendorff@ucl.ac.uk (A.E. Teschendorff).

https://doi.org/10.1016/j.ebiom.2018.04.025

cancers (of which 42 were matched to the normal-adjacent ones), all profiled with Illumina 450 k beadarrays [24]. Since cell type heterogeneity represents a major source of DNAm variation in a complex tissue such as breast, we use recent state-of-the art statistical techniques to rigorously adjust for this major confounder. Using these techniques, as well as an independent validation, we demonstrate that DNAm changes in normal cells are more predictive of breast cancer status than their CNV counterparts.

2. Materials and Methods

2.1. Breast Cancer DNA Methylation Datasets

We analyzed 2 different normal breast and breast cancer tissue datasets, both profiled with the same Illumina Infinium 450 k DNAm technology. The Erlangen set was generated and analyzed by us previously [24], consisting of 50 normal healthy breast samples, 42 age-matched normal-adjacent breast cancer pairs (84 samples in total), and an additional 263 unmatched breast cancers. The clinical characteristics and normalization of the DNAm dataset was described previously [24]. The second "validation" dataset generated Illumina 450 k profiles for 18 normal healthy (from breast reduction surgery) breast samples, as well as 70 normal samples found adjacent to breast cancer [31]. Clinical characteristics and normalization of the Infinium data was described by us previously [24,31].

2.2. Construction and Validation of a Reference DNA Methylation Database for Breast Tissue

We aimed to build a reference DNAm database for breast tissue that would allow us to estimate fractions of epithelial, adipocyte and immune-cells from the DNAm profile of a sample, using the EpiDISH algorithm [32]. To construct the reference database, we used 450 k data representing human mammary epithelial cells (HMECs) from Lowe et al. [33], adipocytes from Nazor et al. [34] and all 7 major immune cell types (neutrophils, eosinophils, monocytes, CD4+ and CD8+ T-cells, B-cells and NK-cells) from Reinius et al. [35]. These 450 k profiles were used in conjunction with an empirical Bayes framework [36] to select differentially methylated CpGs (DMCs) between all 9 cell types, demanding FDR < 0.05 and at least 50% difference in average DNAm between cell types. Cell type specific DMCs were filtered further by demanding that they map to a DNase Hypersensitive Site (DHS), as determined by the NIH Epigenomics Roadmap (if such cell type specific DHS data were available), following a procedure we used previously [32]. This resulted in a reference matrix of 349 DMCs and 9 nine cell types. For an independent sample, cell type fractions for the 9 cell types can be estimated using EpiDISH (using the implementation with Robust Partial Correlations).

We performed three separate validations/tests to ensure that the reference DNAm profiles are representative of epithelial, fat and immune-cells. First, we collected Illumina 450 k data representing these same cell types from independent studies: HMECs from ENCODE [37], adipocytes and blood samples from Slieker et al. and [38]. We constructed 100 in-silico random mixtures of these 3 cell types and compared estimated to true cell-fractions. Second, we applied the reference DNAm profile database and EpiDISH to purified monocytes, T-cells and B-cells from 50 monozygotic twin pairs [39], as this should correctly predict zero fractions for epithelial and adipocytes and near 100% for blood cell types. Third, we applied the reference DNAm profile database and EpiDISH to WGBS data of two IHEC samples enriched for breast epithelial cells [40], as this should predict higher cell-fractions for the epithelial component.

2.3. Identification of DNAm Field Defects

The procedure used to identify epigenetic field defects in normal-adjacent breast tissue was described by us previously [24]. Briefly, we used our iEVORA algorithm to identify differentially variable (DV) and differentially methylated CpGs (DVMCs) between the 50 normal healthy and 42 normal-adjacent samples. The iEVORA algorithm demands genome-wide significance (after correction for multiple testing) at the level of differential variance only, thus defining differentially variable CpGs (DVCs), but subsequently re-ranks DVCs by a t-statistic, in order to favor DVCs where the differential variance is driven by as many outliers as possible. This re-ranking heuristic achieves a good compromise between sensitivity and the type-1 error rate, as demonstrated by us previously [41]. DVMCs were selected using a FDR threshold of 0.001 for differential variability (*P*-values estimated using Bartlett's DV test, which we stress can also be interpreted as a normality deviation test) and a P-value threshold of 0.05 for the t-statistics. Subsequently, we restrict to hypervariable DVMCs, i.e. the subset exhibiting increased variance in the normal-adjacent samples, as the underlying hypothesis is that samples exhibiting deviations from the normal-state represent those at higher risk of carcinogenic transformation.

An appealing feature of using differential variability statistics to identify DNAm alterations in normal-adjacent samples compared to healthy normals is that the resulting hyperV DVMCs are less likely to be driven by changes in cell type composition compared to randomly selected set of CpGs. To see this, we note that the use of the differential variability statistic favors CpGs (hyperV DVMCs) that show ultra-stable DNAm profiles across the normal healthy samples (i.e. very small variance), with outliers driving increased variance in the normal-adjacent specimens. The ultra-high stability of DNAm across the normal healthy samples means that these CpGs are not markers of underlying cell types (in breast these are mainly epithelial cells, adipocytes and immune cells), since variations in the adipose, epithelial and immune cell fractions dominate the top components of variation across normal samples [24]. To prove the result formally, we used our EpiDISH algorithm [32] and our reference DNAm database for breast tissue to estimate epithelial, adipose and immune-cell fractions in all 50 normal samples from healthy women, demonstrating that the top PC in a PCA correlated with these fractions. We then derived CpGs correlating significantly with the estimated epithelial and adipose fractions, thus defining "cell type" DMCs (ctDMCs). We then compared how the previously selected hyperV DVMCs ranked among the list of ctDMCs (i.e those CpGs correlating most strongly with cell type composition) to demonstrate that hyperV DVMCs are ranked significantly lower than a randomly selected set of 10,000 non-DVMCs. We also compared the ranking of the hyperV DVMCs to all non-DVMCs, which did not alter the conclusions.

2.4. CNV Calling Procedure

We used the following procedure to derive copy number alterations for both the Erlangen and validation Illumina 450 k sets. First, idat files were loaded, background-corrected and normalized using functions implemented in the *minfi* package [42]. The returned MethylSet object was then used as input to the *conumee* package [43], to infer CNV states. Briefly, *conumee* performs the inference in 3-steps: (i) background corrected intensity values of the "methylated" and "unmethylated" channels are added, and the log2-ratio of probe intensities of the query sample (this includes any sample, be it normal, normal-adjacent or cancer) to the average over all normal healthy samples is calculated, (ii) the median log2-ratio of probes within predefined genomic bins defines the bin-intensity value, and the bin intensity values are then shifted to minimize the median absolute deviation of all bin intensities from zero to determine the copy-number neutral state, (iii) segmentation is performed using the circular binary segmentation (CBS) algorithm implemented in the *DNAcopy* package [44]. For calling CN gain or loss, we used sample-specific thresholds instead of the widely used cutoffs (± 0.1), in order to reduce the bias caused by cell type heterogeneity. The sample-specific threshold for CN gain/loss is determined automatically by analyzing the distribution of all shifted bin intensity values. For normal-adjacent samples, the median of the log2 ratio

Please cite this article as: Gao, Y., et al., DNA Methylation Patterns in Normal Tissue Correlate more Strongly with Breast Cancer Status than Copy-Number Variants, EBioMedicine (2018), https://doi.org/10.1016/j.ebiom.2018.04.025