Gene 555 (2015) 203–207

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Application of the rank-based method to DNA methylation for cancer diagnosis

Hongdong Li [a], Guini Hong [a], Hui Xu [b], Zheng Guo [a,b,c,*]

[a] Bioinformatics Centre, School of Life Science, University of Electronic Science and Technology of China, Chengdu, China
[b] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China
[c] Department of Bioinformatics, School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China

ARTICLE INFO

Article history:
Received 15 September 2014
Accepted 4 November 2014
Available online 13 November 2014

Keywords:
DNA methylation
Rank-based
Cancer diagnosis

ABSTRACT

Detecting aberrant DNA methylation as diagnostic or prognostic biomarkers for cancer has been a topic of considerable interest recently. However, current classifiers based on absolute methylation values detected from a cohort of samples are typically difficult to be transferable to other cohorts of samples. Here, focusing on relative methylation levels, we employed a modified rank-based method to extract reversal pairs of CpG sites whose relative methylation level orderings differ between disease samples and normal controls for cancer diagnosis. The reversal pairs identified for five cancer types respectively show excellent prediction performance with the accuracy above 95%. Furthermore, when evaluating the reversal pairs identified for one cancer type in an independent cohorts of samples, we found that they could distinguish different subtypes of this cancer or different malignant stages including early stage of this cancer from normal controls. The identified reversal pairs also appear to be specific to cancer type. In conclusion, the reversal pairs detected by the rank-based method could be used for accurate cancer diagnosis, which are transferable to independent cohorts of samples.

1. Introduction

Aberrant DNA methylation in cancer, including global hypomethylation and local hypermethylation of certain genes, is one of the common forms of molecular alterations in carcinogenesis (Baylin et al., 2000). It has been recognized that DNA-based molecular biomarkers, such as DNA methylation patterns, are readily amplifiable and easily translated from a research laboratory setting into routine diagnostics in a clinical trial (Tost, 2010). Therefore, many researchers have tried to detect aberrant DNA methylation changes as diagnostic or prognostic biomarkers for various types of cancer in the past few years (Tost, 2010; Dehan et al., 2009; Levenson, 2010). However, many detected biomarkers usually lack of validation in independent datasets, raising doubt about their transferability. On the other hand, the research efforts towards developing classifiers have often focused on the machine learning methods, such as support vector machine (SVM) (Bhasin et al., 2005) and artificial neural networks (Wang et al., 2010). However, such classifiers are difficult to interpret biological meaning according to the rules of classification and hardly transferable to independent experiments. In recent years, at the transcriptome levels, many studies have successfully applied the relative expression-based method for finding disease biomarkers entirely based on pairs of genes with relative expression values in disease sample reversal to those in the controls (Geman et al., 2004; Tan et al., 2005). Comparing with the machine learning methods, this parameter-free method can avoid data over-fitting and classifiers obtained by this method are biologically interpretable, transferable, and invariable to any monotonic transformation of the data. However, whether this relative ordering of genes could be applied to DNA methylation data for developing disease diagnosis or prognostic biomarkers has not yet been evaluated.

In this study, based on DNA methylation profiles collected from the Cancer Genome Atlas (TCGA database http://tcga-data.nci.nih.gov/tcga), we employed the rank-based methods to detect the relative methylation level reversal pairs (R-pairs) between normal and tumor tissue samples as candidate marker pairs. Then, we identified the most discriminatory R-pairs by considering the top CpG sites with the highest appearance frequencies in all candidate marker pairs. For simplicity, we only focused on the top 11 R-pairs (FR-pairs) which involved 11 CpG sites with the highest appearance frequencies in all candidate marker pairs. The FR-pairs identified for each cancer type performed well in testing sets for the same cancer types and also in validation sets from

Abbreviations: SVM, support vector machine; KNN, k-nearest neighbor; TSP, Top scoring pair; R-pairs, the relative methylation level reversal pairs; FR-pairs, the relative methylation level reversal pairs with frequency; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; COAD, colon adenocarcinoma; STAD, stomach adenocarcinoma; BRCA, breast invasive carcinoma.

* Corresponding author at: Bioinformatics Centre, School of Life Science, University of Electronic Science and Technology of China, No. 4 Section 2, North Jianshe Road, Chengdu 610054, China.
E-mail address: guoz@ems.hrbmu.edu.cn (Z. Guo).

http://dx.doi.org/10.1016/j.gene.2014.11.004
0378-1119/

other independent experiments. Moreover, the FR-pairs also performed well in distinguishing samples with different subtypes and different malignant degrees of the same cancer type they identified from normal controls. Validation on DNA methylation profiles from different cancer types showed that the FR-pairs were specific to cancer type.

2. Materials and methods

2.1. Datasets

The DNA methylation profiles analyzed in this study were downloaded from TCGA database (Cancer Genome Atlas Research Network, 2008) and the Gene Expression Omnibus (GEO) (Barrett et al., 2009). Detailed dataset information was described in Table 1. Each of the five cancer datasets downloaded from TCGA, namely lung adenocarcinoma (LUAD), kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD), stomach adenocarcinoma (STAD) and breast invasive carcinoma (BRCA), was divided into two subsets according to the batch ID provided by the TCGA database: the batch comprising of normal and cancer samples with the largest normal and cancer sample sizes as training set and the remaining batches as testing set (Table 2). Apart from these five cancer datasets, the remaining datasets were used as validation sets.

All profiles were generated using the Human Methylation27 Bead Array (San Diego, CA, USA), targeting 27,578 CpG sites located in promoter regions of unique 14,495 genes. For the datasets collected from TCGA, only level 2 data were used, which included methylated signal intensity (M) and unmethylated signal intensity (U) for each probe. For each CpG site, the methylation level, denoted as a beta-value (β), was calculated as below (Bibikova and Fan, 2009):

$$\beta = \frac{\max(M, 0)}{U + M + 100}. \quad (1)$$

2.2. Detection of R-pairs

For each training set of the five cancer types, we determined the relative methylation level reversal pair (R-pair) by a modified TSP (top scoring pair) method (Geman et al., 2004). For a given dataset, the methylation profiles can be represented as a matrix A with dimension $M \times N$, where M represents the number of CpG sites and N represents the number of profiles. A profile either belongs to *class*1 (normal samples) or *class*2 (tumor samples) and could be denoted as $[\beta_1 \ldots, \beta_i, \ldots, \beta_M]$, where β_i represents the methylation level for CpG site i. If the methylation levels of two CpG sites, k and j, satisfied that the probability of $\beta_k < \beta_j$ in *class*1 significantly differed from that in *class*2, these two CpG sites can be considered as R-pair.

Suppose there are $N1$ samples in *class*1 and $N2$ samples in *class*2 ($N1 + N2 = N$). For a R-pair (k, j), if $\beta_k < \beta_j$ was observed in a samples in *class*1 and b samples in *class*2, then the difference in probability of $\beta_k < \beta_j$ between *class*1 and *class*2 for the pair (k, j) can be calculated by Eq. (2):

$$\begin{aligned} \Delta P_{kj} &= \left| P\left(\beta_k < \beta_j | class1\right) - P\left(\beta_k < \beta_j | class2\right) \right| \\ &= \left| P_{kj}(class1) - P_{kj}(class2) \right| \approx \left| \frac{a}{N1} - \frac{b}{N2} \right|. \end{aligned} \quad (2)$$

For each pair (k, j), another score was used to measure the average rank difference ($\Delta avgR$) between CpG site k and site j from *class*1 to *class*2, which was calculated by Eq. (3):

$$\Delta avgR_{kj} = \left| \frac{\sum_{n=1}^{N1} \left(R_{n,k} - R_{n,j} \right)}{N1} - \frac{\sum_{m=1}^{N2} \left(R_{m,k} - R_{m,j} \right)}{N2} \right| \quad (3)$$

where $N1$ and $N2$ represent the number of profiles in *class*1 and *class*2, respectively. $R_{n,k}$, $R_{n,k}$, $R_{m,i}$, and $R_{m,j}$ represent the rank of site k (or j) in the n-th and m-th profiles of *class*1 and *class*2 respectively.

2.3. Selection of R-pairs as markers

In the process of selecting marker R-pairs for each type of cancer, we first selected the K CpG sites with the highest appearance frequencies in all R-pairs. Then, for each of the K CpG sites, a CpG site was selected and paired to obtain a R-pair according to the following rules: for a CpG site j in K CpG sites, a site i was selected if the Pair (i, j) had the maximum $\Delta avgR$ score among all possible pairs composed of site j. If site i was in K CpG sites or had already been selected by other CpG sites in K CpG sites, then the pair (i, j) is deleted from all possible pairs and the site i is selected according to the rules again.

Table 2
Dataset used in detecting methylation pattern biomarker.

Cancer type[a]	Training set		Testing set		Training set batch ID
	Normal	Tumor	Normal	Tumor	
LUAD	17	35	7	92	58
KIRC	50	50	149	169	64
COAD	11	13	26	164	66
STAD	45	45	12	35	48
BRCA	20	44	7	271	93

[a] LUAD represents lung adenocarcinoma; KIRC represents kidney renal clear cell carcinoma; COAD represents colon adenocarcinoma; STAD represents stomach adenocarcinoma; BRCA represents breast invasive carcinoma.

Table 1
Methylation datasets used in this study.

Cancer type	Abbreviation	Sample size		Data source	Ref.
		Normal	Tumor		
Lung adenocarcinoma	LUAD	24	127	TCGA	Cancer Genome Atlas Research Network (2008)
Kidney renal clear cell carcinoma	KIRC	199	219	TCGA	Cancer Genome Atlas Research Network (2008)
Colon adenocarcinoma	COAD	37	167	TCGA	Cancer Genome Atlas Research Network (2008)
Stomach adenocarcinoma	STAD	57	80	TCGA	Cancer Genome Atlas Research Network (2008)
Breast invasive carcinoma	BRCA	27	315	TCGA	Cancer Genome Atlas Research Network (2008)
Lung squamous cell carcinoma	LUSC	27	133	TCGA	Cancer Genome Atlas Research Network (2008)
Kidney renal papillary cell carcinoma	KIRP	5	16	TCGA	Cancer Genome Atlas Research Network (2008)
Colorectal cancer	CRC44	22	22	GEO (GSE17648)	Kim et al. (2011)
Gastric cancer	GAC75	32	43	GEO (GSE25869)	Kwon et al. (2011)
Breast cancer	BRC248	12	236	GEO (GSE20713)	Dedeurwaerder et al. (2011)