Non-Parametric Change-Point Method for Differential Gene Expression Detection

Yao Wang¹, Chunguo Wu^{1,3,4}, Zhaohua Ji^{1,2}, Binghong Wang^{3,4}, Yanchun Liang^{1*}

1 Key Laboratory for Symbol Computation and Knowledge Engineering of National Education Ministry, College of Computer Science and Technology, Jilin University, Jilin, China, **2** Inner Mongolia Xing'an Vocational & Technical College, Wulanhaote, China, **3** Business School, University of Shanghai for Science and Technology, Shanghai, China, **4** Department of Modern Physics, University of Science and Technology of China, Hefei, China

Abstract

Background: We proposed a non-parametric method, named Non-Parametric Change Point Statistic (NPCPS for short), by using a single equation for detecting differential gene expression (DGE) in microarray data. NPCPS is based on the change point theory to provide effective DGE detecting ability.

Methodology: NPCPS used the data distribution of the normal samples as input, and detects DGE in the cancer samples by locating the change point of gene expression profile. An estimate of the change point position generated by NPCPS enables the identification of the samples containing DGE. Monte Carlo simulation and ROC study were applied to examine the detecting accuracy of NPCPS, and the experiment on real microarray data of breast cancer was carried out to compare NPCPS with other methods.

Conclusions: Simulation study indicated that NPCPS was more effective for detecting DGE in cancer subset compared with five parametric methods and one non-parametric method. When there were more than 8 cancer samples containing DGE, the type I error of NPCPS was below 0.01. Experiment results showed both good accuracy and reliability of NPCPS. Out of the 30 top genes ranked by using NPCPS, 16 genes were reported as relevant to cancer. Correlations between the detecting result of NPCPS and the compared methods were less than 0.05, while between the other methods the values were from 0.20 to 0.84. This indicates that NPCPS is working on different features and thus provides DGE identification from a distinct perspective comparing with the other mean or median based methods.

Citation: Wang Y, Wu C, Ji Z, Wang B, Liang Y (2011) Non-Parametric Change-Point Method for Differential Gene Expression Detection. PLoS ONE 6(5): e20060. doi:10.1371/journal.pone.0020060

Editor: Syed A. Aziz, Health Canada, Canada

Received: December 31, 2010; **Accepted:** April 14, 2011; **Published:** May 31, 2011

Copyright: © 2011 Wang 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 supported by funds from National Natural Science Foundation of China (NSFC) (61073075, 60803052, 10872077, and 91024026), <http://www.nsf.gov.cn/Portal0/default106.htm>; the National High-Tech R&D Program of China (2009AA02Z307), <http://www.863.gov.cn/>; Jilin University ("985" project, "211" project, Scientific Frontier and Interdisciplinary Subject Project (200903173)); the Science-Technology Research Project of Colleges and Universities of Inner Mongolia Autonomous Region of China (NJ10320). 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: ycliang@jlu.edu.cn

Introduction

When normal gene expression is exposed to radiation, virus infection, etc., it would cause gene mutation or gene abnormal activation, which probably leads to cancer arising [1]. There are observable differences between cancer and normal tissues in their expression values on single-gene level, which enables recognition of cancer-related gene from a statistical perspective.

Based on microarray gene expression profiling [2], many methods were reported aiming to detect such difference in gene expression, or normally called differential gene expression (DGE) [3,4]. Among these methods, T-statistics is a classical and widely-used DGE detecting methods, which works on the hypothesis that all the cancer samples are over-expressed compared with the normal samples [5]. Other work has also presented meaningful results, such as empirical Bayes approach [6] (Efron 2001), mixture model approach [7] (Pan, 2003), and SAM [8] (Storey 2003). However, considering the heterogeneity of gene activation, it is reasonable to assume that DGE could only take place in a

subset of cancer samples. Many methods were proposed to solve DGE detection under this assumption, such as PPST (permutation percentile separability test) [9] (Lyons-Weiler, 2004), COPA (cancer outlier profile analysis) [10,11], OS (outlier sum) [12] (Tibshirani, 2007), ORT (outlier robust t-statistics) [13] (Wu, 2007), and MOST (maximum ordered subset t-statistics) [14] (Lian, 2008).

Most of the aforementioned methods attempt to identify the abnormal data points based on the overall percentile of the gene expression profile. However, it is reasonable to assume that the DGE detection could be achieved by searching for the change point of the gene expression profile. If we consider the single-gene expression profile as a data sequence, for non-DGE sequence, there is no significant change between the data distributions of normal and cancer samples; for DGE sequence, since the gene expression is over regulated in cancer group, the data distribution of cancer and normal samples become distinctly different, which would result in a significant change point in the sequence of gene expression profiles.

Table 1. FPR of NPCPS in Monte Carlo simulation.

$n_1 = n_2 = 25, \alpha = 0.05, C(\alpha) = 1.358$									
k	1	3	5	7	9	12	15	20	25
FPR	0.5953	0.1524	0.0333	0.0054	0.0007	0.0001	0.0	0.0	0.0
$n_1 = n_2 = 25, \alpha = 0.01, C(\alpha) = 1.628$									
k	1	3	5	7	9	12	15	20	25
FPR	0.84	0.48	0.163	0.04	0.008	0.002	0.0	0.0	0.0
$n_1 = n_2 = 50, \alpha = 0.05, C(\alpha) = 1.358$									
k	1	4	7	9	12	17	22	30	50
FPR	0.5263	0.0541	0.0040	0.0006	0.0	0.0	0.0	0.0	0.0
$n_1 = n_2 = 50, \alpha = 0.01, C(\alpha) = 1.628$									
k	1	4	7	9	12	17	22	30	50
FPR	0.802	0.268	0.042	0.009	0.0	0.0	0.0	0.0	0.0

doi:10.1371/journal.pone.0020060.t001

Change point problem [15] was widely studied in many fields, such as atmospheric and financial analysis. There are also applications of change-point theory to the microarray analysis, for example, a change point detection model for genomic sequences of continuous measurements [16], ARTIVA formalism for topology inference of regulatory network [17], a Bayesian model for DGE patterns of the DosR regulon of Mycobacterium tuberculosis in the timing of gene induction [18]. With respect to DGE analysis, there are BRIDGE (Bayesian robust inference for differential gene expression) for DGE detection in microarrays with small sample sizes [19], and DGE detecting method LRS (likelihood ratio test) [20] (Hu, 2008).

Since a few of the currently available change-point methods deal explicitly with estimation of the number and location of change points, and moreover these methods may be somewhat vulnerable to deviations of model assumptions usually employed [16], we propose a non-parametric statistical method for DGE detection, named as NPCPS (Non-Parametric Change Point Statistics). NPCPS is based on modified Kolmogorov statistic to detect the single-change point in a data sequence [21]. This method compares the data distribution of normal and cancer group to detect the existence of possible change-point in the cancer

group, and to estimate the position of change-points. Besides, as a non-parametric inferential method, NPCPS does not make assumptions about the probability distributions of the variables being assessed, and accordingly, it is not necessary to normalize the microarray data before calculating the test statistic like other parametric methods usually do. As comparison, we tested several percentile-based methods and LRS. BRIDGE was not included as it is originally designed for two-sample problem and application to larger sample size is computationally heavy. NPCPS works comfortably with large-scale dataset, and both simulation and experiment results show that NPCPS is effective for DGE detection.

Methods

Suppose $X = x_1, \dots, x_r, \dots, x_n$ are independent random variables with cumulative distribution function F , and r is the change point of X . Then, for distribution function F_r of $X_r = x_1, \dots, x_r$ and F_{n-r} of $X_{n-r} = x_{r+1}, \dots, x_n$, there exists a value x_0 that satisfies $F_r(x_0) \neq F_{n-r}(x_0), x_0 \in [x_1, x_n]$. Otherwise, if r is not the change point, we have $F(x_0) = F_r(x_0) = F_{n-r}(x_0), x_0 \in [x_1, x_n]$. The change point is also noted as

Table 2. Actual and average estimate of change point using NPCPS in Monte Carlo simulation.

$n_1 = n_2 = 25, \alpha = 0.01, C(\alpha) = 1.628$									
k	1	3	5	7	9	12	15	20	25
t_0	0.98	0.94	0.90	0.86	0.82	0.76	0.70	0.60	0.50
$\hat{t}_0$	0.6	0.80	0.84	0.82	0.81	0.74	0.689	0.597	0.505
$n_1 = n_2 = 50, \alpha = 0.01, C(\alpha) = 1.628$									
k	1	4	7	9	12	17	22	30	50
t_0	0.99	0.96	0.93	0.91	0.88	0.83	0.78	0.65	0.50
$\hat{t}_0$	0.623	0.880	0.898	0.89	0.868	0.824	0.775	0.650	0.505

t_0 : Actual change point. $\hat{t}_0$: Average estimate of change point.

doi:10.1371/journal.pone.0020060.t002

$$t_0 = r/n, (1 < r < n). \quad (1)$$

Gene expression profile of a single gene could also be considered as a sequence of independent variables as below:

$$\begin{aligned} X_1 &= x_1, \dots, x_{n_1}, \\ X_2 &= x_1, \dots, x_{n_2}, \\ X &= X_1 + X_2 = x_1, \dots, x_{n_1}, \dots, x_n, n = n_1 + n_2. \end{aligned} \quad (2)$$

Here, X_1 contains expression values of normal samples in known distribution function $F_1(x)$, and X_2 contains expression values of disease samples. Over or under expression values in X_2 would result in a change point in X .

To detect the change point, in the hypothesis test we used a modified Kolmogorov statistic (K-statistic), which evaluates the distance between two distribution functions:

$$D_{nk}(x) \triangleq \frac{n-k}{\sqrt{n-k}} (\hat{F}_{(n-k)}(x) - F_1(x)), k=0, 1, \dots, n-1, \quad (3)$$

$\hat{F}_{n-k}(x)$ is the empirical distribution function of $X_{n-k}, \dots, X_n$, defined as:

$$\hat{F}_{n-k}(x) \triangleq \frac{1}{n-k} \sum_{i=k+1}^n I(X_i \leq x), k=0, 1, \dots, n-1, \quad (4)$$

where I is an indicator function. $F_1^{-1}(y)$ is the inverse function of $F_1(x)$ defined as

$$F_1^{-1}(y) \triangleq \inf \{x : F_1(x) \geq y\}, 0 < y < 1, \quad (5)$$

where y is a variable increasing with a fixed step that is subject to user's selection (we selected 100 in the simulation study and experiment).

Therefore, the testing procedure is defined as

$$\begin{aligned} D_n(t, y) \triangleq \frac{1}{\sqrt{n - [nt]}} \sum_{i=[nt]+1}^n \{I(X_i \leq F_1^{-1}(y)) - y\}, \\ 0 < t < 1, 0 < y < 1, \end{aligned} \quad (6)$$

where $[\cdot]$ means round toward negative infinity.

Null hypothesis H_0 is true when $\sup_{0 < t < 1} \sup_{0 < y < 1} |D_n(t, y)| < C(\alpha)$, i.e. no change point detected; alternative hypothesis H_1 is true when $\sup_{0 < t < 1} \sup_{0 < y < 1} |D_n(t, y)| \geq C(\alpha)$, i.e. X has a change point. $C(\alpha)$ is the critical value and α is the significance level. Typical values of $C(\alpha)$ include $C(0.05) = 1.358$ and $C(0.01) = 1.628$.

To give an estimate of change point, we define

$$D_n(t) \triangleq \sup_{0 < y < 1} |D_n(t, y)|, \quad (7)$$

and

$$D_n \triangleq \sup_{0 < t < 1} |D_n(t)|. \quad (8)$$

Let $\hat{t}_0(n)$ be the estimate of t_0 , which is defined as:

$$\hat{t}_0(n) \triangleq \inf \{t : D_n(t) = D_n\}. \quad (9)$$

Since the test statistic measures the difference between two distribution functions, larger $|D_n(t, y)|$ indicates more significant DGE, while the positive $D_n(t, y)$ corresponds to under-expression and the negative $D_n(t, y)$ corresponds to over-expression, respectively.

Results and Discussion

Methods compared with NPCPS

Gene expression profile obtained from microarray data is often considered as a $g \times n$ matrix, which contains g rows of genes with their expression levels in n samples, in which normal group has n_1 samples, and disease group has n_2 samples. Let x_{ij} be the expression intensity of the i th gene in the j th sample of the normal group, while $i = 1, 2, \dots, g, j = 1, 2, \dots, n_1$; let y_{ij} be the expression intensity of the i th gene in the j th sample of the disease group, while $i = 1, 2, \dots, g, j = 1, 2, \dots, n_2$.

Figure 1. FPR and estimate of change-point position. (A) Monte Carlo simulation results of dataset with size $n_1 = n_2 = 25$ and significance level $\alpha = 0.01$. (B) Monte Carlo simulation results of dataset with size $n_1 = n_2 = 50$ and significance level $\alpha = 0.01$. The x-axis is k , the number of samples in simulated dataset that contained DGE. The trend of curves in (A) and (B) was similar. Both FPR and estimate of change-point enhanced with the increasing k . When $k > 9$, the difference between the true change-point and the estimated change-point was very small, and the FPR of NPCPS became lower than the significance level α , which indicated that the hypothesis test of NPCPS passed the Monte Carlo simulation.

doi:10.1371/journal.pone.0020060.g001

..., g , $j = 1, 2, \dots, n_2$. The median of the i th gene is defined as

$$med_i = median(\{x_{ij}\}_{1 \leq j \leq n_1}, \{y_{ij}\}_{1 \leq j \leq n_2}), \quad (10)$$

Define med_{ix} , the normal-group median of the i th gene, and med_{iy} the cancer-group median as

$$med_{ix} = median_{1 \leq j \leq n_1}(x_{ij}), \quad (11)$$

$$med_{iy} = median_{1 \leq j \leq n_2}(y_{ij}). \quad (12)$$

Parametric methods for DGE in cancer subset. There have been many parametric methods proposed based on the mean, median and median absolute deviation (MAD) of the gene expression profile, and following are some typical methods.

COPA [10,11]: COPA first normalizes the expression data using the group mean and MAD to prevent impact to the data distribution hypothesis by outliers, then sorts the expression value

and detects cancer genes through the r th percentile of the cancer group. If the MAD of the i th gene is approximated as

$$mad_i = 1.4826 \times median(\{x_{ij} - med_i\}_{1 \leq j \leq n_1}, \{y_{ij} - med_i\}_{1 \leq j \leq n_2}). \quad (13)$$

The COPA statistic is defined as

$$copa_i = \frac{q_r(\{y_{ij}\}_{1 \leq j \leq n_2}) - med_i}{mad_i}, \quad (14)$$

where $q_r(i)$ is the r th percentile of the i th gene expression values, which is subject to the user's selection.

OS [12]: OS introduced heuristic rule as an additional function, and also applied the percentile knowledge to detect DGE. OS normalizes every gene to ensure the same data scale, which is convenient for gene comparison. In OS, gene expression values greater than $Q_3(i) + IQR(i)$ and smaller than $Q_1(i) - IQR(i)$ are statistically considered as DGE. The OS statistic is defined as below.

Figure 2. Selected ROC curves of normal dataset with $\mu = 2$. (A) $n_1 = n_2 = 25$, $k = 3$. (B) $n_1 = n_2 = 25$, $k = 5$. (C) $n_1 = n_2 = 25$, $k = 9$. (D) $n_1 = n_2 = 50$, $k = 1$. (E) $n_1 = n_2 = 50$, $k = 4$. (F) $n_1 = n_2 = 50$, $k = 9$. The x-axis is FPR, and the y-axis is TPR. The significance level $\alpha = 0.01$ for NPCPS. Larger area under ROC curves indicates better sensitivity and specificity. An ROC curve along the diagonal line indicates random-guess.

doi:10.1371/journal.pone.0020060.g002

$$os_i = \frac{q_r(\{y_{ij}\}_{1 \leq j \leq n_2}) - med_i}{mad_i} \quad (15)$$

The group with over expression is defined as

$$W_i = \sum_{j=1}^{n_2} y_{ij} I\{y_{ij} > q_{75}(i) + IQR(i)\} \quad (16)$$

Similarly, the group with under expression DGE is defined as

$$W_i = \sum_{j=1}^{n_2} y_{ij} I\{y_{ij} < q_{25}(i) - IQR(i)\} \quad (17)$$

ORT [13]: Compared with OS, ORT is similar but uses the median of the normal group instead of the median of all data, and estimates the absolute error using the median of

several groups instead of the square error as in COPA [11]. The purpose of these changes is to acquire more robust and consistent estimation. Accordingly, the estimate MAD in ORT is

$$mad_{ij} = 1.4286 \times median\{|x_{ij} - med_{ix}|_{1 \leq j \leq n_1}, |y_{ij} - med_{iy}|_{1 \leq j \leq n_2}\} \quad (18)$$

The ORT statistic is

$$ort_j = \frac{\sum i \in C_j (\{y_{ij}\}_{1 \leq j \leq n_2} - med_{ix})}{mad_{ij}}, \quad (19)$$

where C_i is the cancer group of the i th gene. For over expression,

$$C_i = \sum_{j=1}^{n_2} y_{ij} I\{y_{ij} > q_{75}(\{x_{ij}\}_{1 \leq j \leq n_1}) + IQR(\{x_{ij}\}_{1 \leq j \leq n_1})\}. \quad (20)$$

Figure 3. Selected ROC curves of normal dataset with $\mu = 1$. (A) $n_1 = n_2 = 25$, $k = 6$. (B) $n_1 = n_2 = 25$, $k = 9$. (C) $n_1 = n_2 = 25$, $k = 14$. (D) $n_1 = n_2 = 50$, $k = 6$. (E) $n_1 = n_2 = 50$, $k = 9$. (F) $n_1 = n_2 = 50$, $k = 15$. The x-axis is FPR, and the y-axis is TPR. The significance level $\alpha = 0.01$ for NPCPS. Larger area under ROC curves indicates better sensitivity and specificity. An ROC curve along the diagonal line indicates random-guess.
doi:10.1371/journal.pone.0020060.g003

Figure 4. Selected ROC curves of skew-normal dataset. (A) $n_1 = n_2 = 25$, $\mu = 2$, $k = 3$. (B) $n_1 = n_2 = 25$, $\mu = 2$, $k = 5$. (C) $n_1 = n_2 = 25$, $\mu = 2$, $k = 9$. (D) $n_1 = n_2 = 50$, $\mu = 2$, $k = 4$. (E) $n_1 = n_2 = 50$, $\mu = 2$, $k = 9$. (F) $n_1 = n_2 = 25$, $\mu = 1$, $k = 6$. (G) $n_1 = n_2 = 25$, $\mu = 1$, $k = 9$. (H) $n_1 = n_2 = 50$, $\mu = 1$, $k = 9$. (I) $n_1 = n_2 = 50$, $\mu = 1$, $k = 14$. The x-axis is FPR, and the y-axis is TPR. The significance level $\alpha = 0.01$ for NPCPS. Larger area under ROC curves indicates better sensitivity and specificity. An ROC curve along the diagonal line indicates random-guess.
doi:10.1371/journal.pone.0020060.g004

Similarly, for under expression,

$$C_i = \sum_{j=1}^{n_2} y_{ij} I\{y_{ij} < q_{25}(\{x_{ij}\}_{1 \leq j \leq n_1}) - IQR(\{x_{ij}\}_{1 \leq j \leq n_1})\}. \quad (21)$$

MOST [14]: Gene with expression value greater than $MOST_{ik}$ is considered as a differential gene expression. The testing statistic $MOST_{ik}$ is defined as:

$$MOST_{ik} = \max_{1 \leq k \leq n_1} \left(\frac{\sum_{1 \leq j \leq k} (y_{ij} - \text{median}(\{x_{ij}\}_{1 \leq j \leq n_0}))}{1.4826 \times \text{median}(X'_i \cup Y'_i)} \right). \quad (22)$$

When k is unknown, the data are normalized firstly by μ_k and σ_k^2 , and the $MOST_{ik}$ is defined as

$$MOST_{ik} = \max_{1 \leq k \leq n_1} \left(\frac{\sum_{1 \leq j \leq k} (y_{ij} - \text{median}(\{x_{ij}\}_{1 \leq j \leq n_0}))}{1.4826 \times \text{median}(X'_i \cup Y'_i)} - \mu_k \right) / \sigma_k. \quad (23)$$

Non-parametric methods for DGE in cancer subset. PPST

[9]: As a non-parametric method, PPST compares the expression levels of thousands of genes in two sample groups, i.e. the control group (A) and the case group (B). The detection focuses on genes in group A of which the expression levels are higher than a certain percentile of group B's expression values ($A > B$), which is a type I error in statistics, and vice versa ($B > A$). There are two marks for each gene, s_1 and s_2 . S_1 is the number of samples in group A that are higher than the 95% of group B added by the number of samples in group B that are lower than the 95% of group A. S_2 is defined as the opposite to s_1 . Considering a given gene, if the expression value in group A is higher than the 95th percentile of group B, it is considered as over expression; if the expression value is lower than that in group B, it is considered as under expression. Define the PPST statistic of each gene with over expression as:

$$ppst_i = \sum_{i=1}^m y_{ij} I\{Z_i, y_{ij} > q_{95}(x_{i1}, \dots, x_{in})\} \quad (24)$$

PPST statistic of gene with under expression can be obtained similarly.

LRS [15]: in LRS, cancer outlier samples are viewed as coming from a distribution with higher mean expression intensity than all the normal and other cancer samples. The purpose of LRS is to test such unequal mean. For up-regulation, LRS first organizes all the samples so the non-cancer samples are arranged before the cancer samples, and the cancer samples are sorted by their expression intensities in the ascending order. S_n is the summation of the expression intensities of all the samples and the LRS statistic is as follows,

$$\lambda = \max_{m_0 \leq i \leq m_1} \frac{iS_n/n - S_i}{[i(1-i/n)]^{1/2}}, n_1 + 1 \leq m_0 < m_1 < n. \quad (25)$$

Traditional methods for DGE in entire cancer group. T-statistic

[5]: this traditional method assumes that the cancer sample group is generally over or under-expressed compared with the normal samples. The t-statistic is defined as:

Table 3. AUC of ROC curves of the simulation on data in normal distribution.

Data Parameter			AUC							
n	mu	k	NPCPS	LRS	COPA	OS	ORT	PPST	T	MOST
50	2	3	0.79	0.82	0.72	0.75	0.76	0.69	0.68	0.67
50	2	5	0.90	0.92	0.82	0.80	0.85	0.79	0.79	0.76
50	2	9	0.96	0.97	0.87	0.81	0.92	0.91	0.91	0.89
100	2	1	0.52	0.57	0.53	0.28	0.58	0.54	0.53	0.52
100	2	4	0.85	0.84	0.72	0.81	0.82	0.69	0.69	0.67
100	2	9	0.96	0.96	0.88	0.90	0.92	0.85	0.86	0.84
50	1	6	0.73	0.74	0.63	0.58	0.65	0.68	0.72	0.67
50	1	9	0.82	0.81	0.67	0.61	0.72	0.76	0.79	0.74
50	1	14	0.90	0.89	0.73	0.65	0.80	0.87	0.90	0.87
100	1	6	0.71	0.70	0.60	0.59	0.59	0.66	0.65	0.63
100	1	9	0.79	0.76	0.66	0.64	0.69	0.66	0.70	0.67
100	1	15	0.88	0.85	0.70	0.66	0.75	0.78	0.82	0.80

doi:10.1371/journal.pone.0020060.t003

$$T_i = \frac{\bar{y}_i - \bar{x}_i}{S_i} \quad (26)$$

where $\bar{x}_i$ is the sample mean of normal group expression values and $\bar{y}_i$ is the sample mean of cancer group expression values, S_i is the estimate of combined standard deviation. Differentially expressed genes are recognized when the testing statistic exceeds a certain threshold.

Monte Carlo simulation

Monte Carlo simulation can be used to evaluate the performance of a hypothesis test in terms of the ratio of Type I error, i.e. false positive rate (FPR). For each Monte Carlo simulation, NPCPS was applied to an artificial 7000-gene dataset (normal random numbers with mean = 0, standard deviation sd = 1) composed of n_1 normal samples and n_2 cancer samples, of which k ($0 < k < n_2$) cancer samples contained DGE simulated by adding a

Table 4. AUC of ROC curves of the simulation on data in skew-normal distribution.

Data Parameter			AUC							
n	mu	k	NPCPS	LRS	COPA	OS	ORT	PPST	T	MOST
50	2	3	0.73	0.69	0.58	0.57	0.65	0.62	0.64	0.61
50	2	5	0.87	0.80	0.61	0.59	0.69	0.69	0.75	0.71
50	2	9	0.95	0.91	0.66	0.59	0.80	0.82	0.90	0.83
100	2	4	0.78	0.71	0.58	0.58	0.63	0.63	0.66	0.63
100	2	9	0.95	0.85	0.64	0.64	0.74	0.72	0.80	0.75
50	1	6	0.66	0.60	0.54	0.55	0.59	0.63	0.66	0.63
50	1	9	0.75	0.70	0.60	0.56	0.63	0.70	0.76	0.71
100	1	9	0.72	0.64	0.57	0.57	0.62	0.63	0.70	0.67
100	1	14	0.82	0.71	0.60	0.58	0.66	0.69	0.78	0.73

doi:10.1371/journal.pone.0020060.t004

Figure 5. Change-point distribution of DGE genes when $C(0.05) = 1.628$. (A) CP-position distribution of 989 genes with positive $D_n > 1.628$. (B) CP-position distribution of 989 genes with negative $D_n < -1.628$.
doi:10.1371/journal.pone.0020060.g005

Table 5. Results and description of top-ranked genes with positive test statistic D_n .

NPCPS Ranking	D_n	CP Position	Gene Name	Description
1	2.91	27	PDE4B	The phosphodiesterase PDE4B limits cAMP-associated PI3K/AKT-dependent apoptosis in diffuse large B-cell lymphoma.
2	2.80	34	N/A	
3	2.79	32	N/A	
4	2.74	27	SCARB2	
5	2.74	27	C9	Upregulation of plasma C9 protein in gastric cancer patients.
6	2.73	22	RAB2A	
7	2.73	27	MEF2A	
8	2.69	37	N/A	
9	2.66	19	ITK	ITK-SYK causes a T-cell lymphoproliferative disease in mice, supporting its role in T-cell lymphoma development in humans.
10	2.66	41	TCF3	Misregulation plays a role in disease such as cancer, where overactive Wnt signaling drives LEF/TCFs to transform cells
11	2.64	31	JAG1	Heterogeneity of Jagged1 expression in human and mouse intestinal tumors: implications for targeting Notch signaling
12	2.64	16	HMGA2	LEF/TCFs to transform cells
13	2.62	28	NEFL	
14	2.60	16	IL2	The development of breast tumor is associated with an increased expression of IL-2 and this expression also seems to be associated with the malignancy of the tumor.
15	2.60	26	RARRES1	RARRES1 expression is significantly related to tumor differentiation and staging in colorectal adenocarcinoma
16	2.60	15	PRKCB	Target for inhibiting gastric cancer cell invasion
17	2.58	14	PRPS2	
18	2.56	19	IFNA10	

doi:10.1371/journal.pone.0020060.t005

Table 6. Results and description of top-ranked genes with negative test statistic D_n .

NPCPS Ranking	D_n	CP Position	Gene Name	Description
5336	-3.08	27	AGER	Serum sRAGE levels were influenced by genetic polymorphisms (-429 T/C, Gly82Ser and 2184 A/G) of the RAGE gene in breast cancer
1929	-2.91	27	GP1BB	
5918	-2.87	27		
2064	-2.87	27	MAPK14	The expression of p-p38 and uPA was negatively correlated to prognosis of breast cancer.
4931	-2.73	26	SLC5A5	The findings of this study indicated that NIS expression is prevalent in breast cancer brain metastases and could have a therapeutic role via the delivery of radioactive iodide and selective ablation of tumor cells
4634	-2.72	27	BMP1	
2124	-2.71	27		
5753	-2.69	27	MYOG	
5907	-2.62	32	MMP11	MMP-11 and CK-20 are probable prognostic markers whose expression reflects the stages of tumor differentiation and LNM of breast cancer
853	-2.60	43	NCSTN	NCSTN coded protein is a subunit of γ -Secretase compound, which is related to Notch signaling, a pathway found dysregulated in many cancers.
4869	-2.54	30	SLC4A2	
5060	-2.53	25	NAT6	
4128	-2.53	27	ALDH4A1	
4516	-2.52	21	SNAPC1	
2257	-2.51	27	WRN	The variant genotype of WRN Leu1074Phe was associated with a 1.36-fold significantly increased risk of breast cancer in Chinese women. This variant is also significantly associated with age at menarche
6187	-2.51	27	ENG	Elevated pretreatment plasma endoglin levels predicted for decreased clinical benefit and a shorter overall survival in metastatic breast cancer patients treated with 2nd-line hormone therapy.
2183	-2.51	27	TRADD	TRADD is involved in the p75(NTR)-mediated antiapoptotic activity of nerve growth factor in breast cancer cells
4444	-2.51	27	UNC119	

doi:10.1371/journal.pone.0020060.t006

constant μ to the original normal random numbers. Multiple simulations were carried out according to different values of sample size n , DGE sample size k , and significance level α . The FPR (Table 1) and average estimate of change point (Table 2) were computed and the results of simulation with $\alpha = 0.01$ were illustrated in Fig. 1. For data set $n_1 = n_2 = 25$ (Fig. 1A), the FPR was larger when k was smaller; FPR decreased with k increasing; when k was equal to or larger than 9, the detecting accuracy of NPCPS was sufficient to satisfy the significance level. For data set $n_1 = n_2 = 50$ (Fig. 1B), k should be not less than 9 to satisfy the significance level. The estimate of change point enhanced greatly when k increased; the estimated position became very close to the actual position at the same time as the FPR dropped below the significance level. This indicates that NPCPS is highly sensitive to left boundary and less sensitive to the right boundary, when the F_2 information is not sufficient [21].

ROC analysis on simulated data

First, we test NPCPS ($\alpha = 0.01$) and seven other methods, namely COPA, ORT, OS, MOST, T, LRS, and PPST, on normally distributed datasets (mean = 0, sd = 1) with different μ , n

and k . When k was getting greater, all methods produced better ROC (Fig. 2 and Fig. 3). For $\mu = 2$, when $n = 50$ (Fig. 2A–2C), NPCPS was slightly weaker than LRS, and better than the other methods; when $n = 100$ (Fig. 2D–2E), NPCPS was very similar to LRS, and better than the other methods. For $\mu = 1$, NPCPS gave the best performance for both $n = 50$ and $n = 100$ datasets and different values of k (Fig. 3A–3F). This indicated that NPCPS had better sensitivity for less significant DGE compared with the other seven methods. Among the non-parametric method, PPST was not significantly better than the parametric methods, while LRS and NPCPS were continuously better than the other methods. This indicated that methods based on change-point were more effective and robust than methods based on percentile and MAD.

Second, we tested NPCPS ($\alpha = 0.01$) and other seven methods on datasets generated from skew-normal (SN) distribution (Fig. 4). For different n and k , NPCPS had significantly larger area under the ROC curves compared with the other methods. By comparing Fig. 2 and Fig. 4 we can see that NPCPS was both effective for normal and skew-normal data distribution, and when $k = 9$ gave similarly good ROC (Fig. 2C compared with Fig. 4C, Fig. 2F compared with Fig. 4E). The other seven methods, including

non-parametric methods, LRS and PPST, had inferior results when working with skew-normal data.

The AUC of ROC is summarized in Table 3 and 4.

DGE detection in breast-cancer microarray data

The microarray data used in the experiment are provided by West [22]. In their experiment, primary breast tumors (between 1.5 and 5 cm in maximal dimension) from the Duke Breast Cancer SPORE frozen tissue bank were selected and diagnosed as invasive ductal carcinoma. In each case, a diagnostic axillary lymph node dissection was performed. The final dataset includes 49 samples, 25 samples of which have negative lymph nodes and 24 samples with positive lymph nodes, used here as normal sample and cancer sample, respectively. Gene expression profile of 7129 genes was obtained through annotation package hu6800 [23]. The original gene expression values ranged from 34 to 43053 and were initialized to the range from 3.5 to 10.7. Seven detection method (t-statistic, COPA, OS, ORT, MOST, PPST and LRS) were

applied to the initialized gene expression profile, while NPCPS was applied to the original data. The calculated test statistics of these 7129 genes by these methods were sorted in descending order.

For NPCPS, $C(0.05) = 1.628$ was selected, which yields a detecting result of 1978 DGE genes. Fig. 5 shows the distribution of the estimated position of change-points in the expression value of these genes. We selected the first 30 genes ranked by NPCPS, and searched PubMed and other databases to confirm that whether these genes were relevant to breast cancer or other known cancers. Out of the first 30 genes identified by NPCPS, 17 have been reported as relevant to breast cancer or other cancers (as shown in Table 5 and Table 6 separately according to D_n value). The gene expression values and the change-point (CP) positions of the cancer-relevant genes are illustrated in Fig. 6 and Fig. 7. From Fig. 6 and 7, it could be seen that the estimated change-point positions could successfully locate the change in the trend of the gene expression value.

Figure 6. Expression value and Change-point of top-ranked DGE genes with positive D_n . (A) PDE4B, change-point at sample 27. (B) C9, change-point at sample 27. (C) ITK, change-point at sample 27. (D) TCF3, change-point at sample 41. (E) JAG1, change-point at sample 31. (F) HMGA2, change-point at sample 16. (G) RARRES1, change-point at sample 26. (H) PRKCB, change-point at sample 15. CP position correctly locates the change in the trend of the gene expression value.

doi:10.1371/journal.pone.0020060.g006

Figure 7. Expression value and Change-point of top-ranked DGE genes with negative D_m . (A) AGER, change-point at sample 27. (B) MAPK14, change-point at sample 27. (C) SLC5A5, change-point at sample 26. (D) MMP11, change-point at sample 32. (E) NCSTN, change-point at sample 43. (F) WRN, change-point at sample 27. (G) ENG, change-point at sample 27. (H) TRADD, change-point at sample 27. CP position correctly locates the change in the trend of the gene expression value.
doi:10.1371/journal.pone.0020060.g007

Table 7. Ranking relevance between each DGE detecting methods.

	NPCPS	LRS	COPA	OS	ORT	PPST	T	MOST
NPCPS	1	0.0354	-0.0346	0.0069	-0.0311	-0.0357	-0.0398	-0.0470
LRS	0.0354	1	0.5007	0.5274	0.3206	0.4182	0.4964	0.4474
COPA	-0.0346	0.5007	1	0.5339	0.5175	0.5825	0.5140	0.5752
OS	0.0069	0.5274	0.5339	1	0.4443	0.5407	0.5868	0.4539
ORT	-0.0311	0.3206	0.5175	0.4443	1	0.7921	0.2032	0.3069
PPST	-0.0357	0.4182	0.5825	0.5407	0.7921	1	0.3645	0.4378
T	-0.0398	0.4964	0.5140	0.5868	0.2032	0.3645	1	0.8394
MOST	-0.0470	0.4474	0.5752	0.4539	0.3069	0.4378	0.8394	1

doi:10.1371/journal.pone.0020060.t007

Figure 8. ROC curves of NPCPS and other methods when inappropriate formula applied. (A) Over expression formula applied to simulated under expressed gene, $n_1 = n_2 = 25$, $\mu = -2$, $k = 8$. (B) Under expression formula applied to simulated over expressed gene, $n_1 = n_2 = 25$, $\mu = 2$, $k = 8$. The x-axis is FPR, and the y-axis is TPR. ∇ is T, $\times$ is COPA, $\circ$ is OS, $\bullet$ is ORT, $\diamond$ is MOST, dotted line is LRS, dashed line is PPST, and solid line is NPCPS. The significance level $\alpha = 0.01$ for NPCPS. NPCPS maintained the same level of sensitivity when applied to both types of simulated over-expressions. The other methods were not able to give results as good as when appropriate functions were applied as in Fig. 2 and Fig. 3. doi:10.1371/journal.pone.0020060.g008

Moreover, by comparison among the ranking results of all eight methods based on the test statistic, it was noticed that genes top-ranked by NPCPS were ranked considerably lower by other methods, most of which were mean and median based parametric methods. When inspecting all the 7129 genes, the overall trend in ranking difference between NPCPS and other methods became

more obvious. Table 7 shows the pair-wise linear correlation of gene ranking among the six methods. For NPCPS, the positive correlation is below 0.007 with OS, and the negative correlation below 0.047 with other methods. This indicated that NPCPS had much less correlation with the other five methods, among which the correlations were all positive and valued around 0.5 with each other.

Figure 9. Data distributions of genes top-ranked by NPCPS. (A) IQGAP1: rank 19, positive D_n . (B) PIP5K1B: rank 20, positive D_n . (C) UBB: rank 21, negative D_n . (D) RFC1: rank 22, negative D_n . Top-ranked genes by NPCPS had significant difference between the data distributions of cancer and normal groups. By comparing the empirical distribution of cancer and normal samples, (A) and (B) had distributions of cancer group that were significantly left to the distribution of normal group, which demonstrated under-expression; (C) and (D) had distributions of cancer group that were significantly right to the distribution of normal group, which demonstrated over expression. The distribution curves were consistent with the biological significance of D_n value. doi:10.1371/journal.pone.0020060.g009

This will be further discussed in the following section. If NPCPS is combined with other methods, it would help to identify genes which are considered as less DGE significant by the other seven methods.

Discussions on the biological significance of NPCPS

Non-parametric statistics. As a non-parametric statistics, NPCPS does not rely on assumptions that the data are drawn from a given probability distribution. It is applicable to input data derived from various types distributions and doesn't require data pre-processing. As such it is the opposite of parametric statistics, which would have inferior performance when the input data are not in the assumed distribution, as in the ROC simulation on normal and skew-normal datasets (Fig. 4).

No restriction on both over expression and under expression. The gene expression profile generated from microarray data usually contains samples of thousands of genes. Genes in the cancer samples might be over or under expressed. Majority of the DGE detecting methods have different formulas for under expressed and over expressed genes, respectively. For example, OS and ORT use different percentile values for over-expression and under-expression, respectively, and apply both formulas to the same microarray data. If over expression formula is applied to under expressed data, the DGE can not be correctly recognized. However, the detected results might contains false alarms, since both over-expression and under-expression formulas are applied to the same gene, and might be detected as DGE significant

Figure 10. Data distribution of genes bottom-ranked by NPCPS. (A) SLC6A8: rank 500. (B) HLF: rank 1000. (C) ATP5F1: rank 2000. (D) HLA-H: rank 3000. (E) ODF3B: rank 4000. (F) SLC20A2: rank 5000. (G) SGSH: rank 6000. (H) CXCR2: rank 7000. From the empirical data distribution, the differences between cancer and normal groups in (A)–(D) were very small, which corresponded with the D_n value.
doi:10.1371/journal.pone.0020060.g010

for twice. Unlike the other methods, NPCPS works for both types of DGE by using the same calculating formula, which would reduce the FDR, and do not require further analysis and computation aiming to clean the false alarms. When over-expression formula was applied to under-expressed gene data (Fig. 8A), and vice versa (Fig. 8B), NPCPS presented stable performance in both situations, while other compared methods gave inferior ROC curves. According to the characteristic of ROC, T and MOST could have good ROC if the prediction result was inversed. The ROC curves of LRS were in the zone of random guess, which was close to the line-of-no-discrimination. Using LRS for under-expression, user could turn under-expression into over-expression by inverting the dataset. This indicated that when

over-expression formula of LRS was applied to under-expression, the random detecting result would be given.

Estimated change point position: $\hat{t}_0$

The biological meaning of $\hat{t}_0$ lies in that once the position of change point is estimated or located, we can identify which sample contains DGE. Then, rather than identifying DGE existence in $n = n_1 + n_2$ samples on the single gene level, we can learn that, for one sample containing thousands of genes, how many genes were over expressed or under expressed. This statistical information can be used to analyze features of each sample, and the results of which could be applied to the estimation of the differentiation degree of cancer in different development stages.

Figure 11. Data distributions of genes top-ranked by five parametric methods. (A) Gene TFF1, COPA rank: 8, NPCPS rank: 5281. (B) Gene ERBB2, COPA rank: 18, NPCPS rank: 6256. (C) Gene ZNF44, OS rank: 6, NPCPS rank: 7057. (D) Gene RGS2, OS rank: 24, NPCPS rank: 6222. (E) Gene TCN1, ORT rank: 7, NPCPS rank: 7113. (F) Gene TRAPPC10, ORT rank: 9, NPCPS rank: 6732. (G) Gene CLDN10, T rank: 22, NPCPS rank: 1107. (H) Gene CSNK1E, T rank: 24, NPCPS rank: 1163. (I) Gene COL2A1, MOST rank: 10, NPCPS rank: 7126. (J) Gene SCG2, MOST rank: 10, NPCPS rank: 2585. Top-ranked genes by the five parametric methods did not have significant difference between the data distributions of cancer and normal groups. doi:10.1371/journal.pone.0020060.g011

Distance between two distribution function: D_n

NPCPS results showed that, among the 7219 genes, 3608 had negative D_n , while the rest 3521 had positive D_n . NPCPS use D_n to evaluate the change in distribution between normal and cancer samples, and directly measure the DGE type as either over expressed or under expressed. This feature is valid based on the expression value in Fig. 6 and 7, where Fig. 6 (positive D_n) shows typical under expression and Fig. 7 (negative D_n) shows typical over expression. Fig. 9 and Fig. 10 can illustrate the relationship between D_n and DGE in a more intuitive manner where cumulative data distributions of several typically ranked genes are given.

Genes in Fig. 9 were ranked on the top by NPCPS, where Fig. 9A and 9B are corresponding to positive D_n (under-expression), 9C and 9D are corresponding to negative D_n (over-expression), respectively. By comparing the empirical distribution of cancer and normal samples, in Fig. 9A and 9B, cancer group was significantly left to the normal group, which demonstrated under expression; in Fig. 9C and 9D, the cancer group was significantly right to the normal group, which demonstrated over-expression. The distribution graph was consistent with the D_n value.

Genes in Fig. 10 were ranked lower by NPCPS. We can find that the cumulative distance between the data distributions of normal and cancer group is generally smaller compared with those genes top-ranked by NPCPS. From the empirical data distribution, difference between cancer and normal groups were very small.

As comparison, Fig. 11A–11J shows the data distributions of those top-ranked genes by the parametric methods, and Fig. 12A–12D by LRS and PPST. The data distributions were more similar to genes that were bottom-ranked by NPCPS in that small percent of the samples bring significant increase to data range. These few samples would greatly impact the cancer-group mean or median, which consequently result in a high test statistic of parametric

methods. For example, in Fig. 11B and 12B, 96% of the two curves were close to each other while 4% data points in the normal group valued much greater, which equals to one outlier sample out of the 25 normal samples. Considering that the outliers were in the normal group, it was reasonable to assume that these outliers might be caused by microarray noise. For the rest of Fig. 11 and 12, except for T-statistic, the cancer group had one outlier. Fig. 11 and 12 indicate that the comparing methods are sensitive to significant change in mean and median, even when the change is introduced by a single sample which might be outliers. NPCPS is less prone to report a DGE as such few outliers are not sufficient to produce a large D_n .

In summary, NPCPS is less sensitive to right boundaries and tends to find genes that have greater cumulative distance between the data distribution of normal and cancer groups. For such genes, the samples in normal and cancer group may have the same data range but should have very different distributions. Therefore, the detecting result of NPCPS would be different from other compared methods, which are more sensitive to outliers that influence the data range, rather than the cumulative distance between distributions. In other words, NPCPS values continuous change in data distribution over the whole data range, while the other methods look for a significant change of mean or median. This would explain the low correlation between NPCPS and other methods.

Conclusion

A non-parametric statistical method, NPCPS, was proposed for DGE detection based on change-point theory. NPCPS uses the data distribution of normal and cancer samples as the only input to detect a change point that indicates DGE, in order to identify potential cancer genes. Distribution-based NPCPS does not require data pre-initialization and is computationally efficient compared with other median-based parametric methods. Contrast

Figure 12. Data distributions of genes top-ranked by two non-parametric methods. (A) Gene MGP, LRS rank: 11, NPCPS rank: 7049. (B) Gene IGF2, LRS rank: 21, NPCPS rank: 7094. (C) Gene TNNC1, PPST rank: 21, NPCPS rank: 3697. (D) Gene E4F1, PPST rank: 47, NPCPS rank: 2595. Top-ranked genes by the two non-parametric methods did not have significant difference between the data distributions of cancer and normal groups. doi:10.1371/journal.pone.0020060.g012

to the compared methods, NPCPS deals with both over-expression and under-expression by the same equation. Another unique feature is that the proposed NPCPS could estimate both the number and the location of cancer samples with DGE could be estimated. Simulation study and experiments showed that, the proposed NPCPS method had better reliability and accuracy; NPCPS was more effective than the compared parametric methods; similar ROC curves was given compared with LRS when sample size was larger; when the simulated DGE value was smaller, i.e. DGE was less significant, NPCPS had better sensitivity compared with the other seven methods. Simulations also indicated that, for cancer subgroup with size greater than 8, the NPCPS had FPR less than 0.01. Besides, the detection results of NPCPS had very low correlation with the compared methods,

both parametric and non-parametric, which indicates that NPCPS provides meaningful detection results different from other methods. Since cancer samples could be categorized according to different stages in the cancer development, DGE detection can also be considered also a multi-class problem. Further effort could be focused on the multi-change-point in the distribution of microarray gene expression profile.

Author Contributions

Conceived and designed the experiments: YW ZJ CW YL. Performed the experiments: YW ZJ. Analyzed the data: YW CW ZJ YL. Contributed reagents/materials/analysis tools: YW CW ZJ BW YL. Wrote the paper: YW CW ZJ YL.

References

1. Tibshirani R, Hastie T (2007) Outlier sums for differential gene expression analysis. *Biostatistics* 8: 2–8.
2. Magic Z, Radulovic S, Brankovic-Magic M (2007) cDNA microarrays: identification of gene signatures and their application in clinical practice. *Journal of B.U.ON* 12 Suppl 1: S39–44.
3. Brent R (2000) Genomic biology. *Cell* 100: 169–183.
4. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, et al. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* 24(3): 227–235.
5. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 100: 8418–8423.
6. Efron B, Tibshirani R, Storey J, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 96: 1151–1160.
7. Pan W, Lin J, Le C (2003) A mixture model approach to detecting differentially expressed genes with microarray data. *Funct Integr Genomics* 3(3): 117–124.
8. Storey JD, Tibshirani R (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays, in the *Analysis of Gene Expression Data: Methods and Software*, Parmigiani G, Garrett ES, Irizarry RA, Zeger SL, eds. New York: Springer.
9. Lyons-Weiler J, Patel S, Becich MJ, Godfrey TE (2004) Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics* 5: 110.
10. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in Prostate Cancer. *Science* 10(310): 644–648.
11. MacDonald JW, Ghosh D (2006) COPA-cancer outlier profile analysis. *Bioinformatics* 22: 2950–2951.
12. Tibshirani R, Hastie T (2007) Outlier sums for differential gene expression analysis. *Biostatistics* 8: 2–8.
13. Wu B (2007) Cancer outlier differential gene expression detection. *Biostatistics* 8(3): 566–575.
14. Lian H (2008) MOST: detecting cancer differential gene expression. *Biostatistics* 9(3): 411–418.
15. Basseville M, Nikiforov IV (1993) *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs: Prentice Hall.
16. Muggeo V, Adelfio G (2011) Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics* 27: 161–166.
17. Lèbre S, Becq J, Devaux F, Stumpf M, Lelandais G (2010) Statistical inference of the time-varying structure of gene-regulation networks. *BMC System Biology* 4: 130.
18. Zhang Y, Hatch KA, Wernisch L, Bacon J (2008) A Bayesian change point model for differential gene expression patterns of the DosR regulon of mycobacterium tuberculosis. *BMC Genomics* 9: 87.
19. Gottardo R, Raftery AE, Yeung KY, Bumgarner RE (2006) Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* 62: 10–18.
20. Hu JH (2008) Cancer outlier detection based on likelihood ratio test. *Bioinformatics* 24(19): 2193–2199.
21. Tan ZP, Miao BQ (2000) Nonparametric Statistical Inference for Distribution Change Point Problems. *Journal of China University of Science and Technology* 6(3): 270–276.
22. West M, Blanchette C, Dressman H, Huang E, Ishida S, et al. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 98: 11462–11467.
23. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* vol 5: R80.