Bimodal Gene Prediction via Gap Maximisation

Abdullatif S. Al-Watban[1,2]**, Zheng Rong Yang**[1]
[1]School of Biosciences, University of Exeter, EX4 4QD, UK
[2]Saudi Food and Drug Authority, Medical Devices Sector, Riyadh, KSA

Abstract- *Bimodal gene is one of the common phenomena frequently observed in gene expression data for certain types of studies including cancer studies and drug/therapy effect studies. There have been several algorithms proposed to predict bimodal genes with success. However, occasionally their performance is not very satisfied. We propose a new algorithm to detect bimodal genes. The new algorithm is based on the assumption that the bimodality is related with the gap between two consecutive expressions. We show that this new algorithm demonstrates better performance compared with several benchmark algorithms using both real and simulated data sets.*

Keywords: bimodal distribution, non-parametric analysis, differential genes, heterogeneity.

1 Introduction

Microarray experiments have benefitted the discovery of genetic differentiation pattern for interpreting the observed phenotypic differentiation for a decade [1]. The success is due to high-throughput and genome-wide examination. The discovery of differential genes in relation to phenotypic differentiation can be implemented using standard student *t* test if data satisfy the assumption. However biological diversity makes this difficult because a large number of genes appear to have bimodal or multi-modal distribution [2]. Fig 1 shows such a typical bimodal distribution of samples in the same category (such as cancer samples) of a gene.

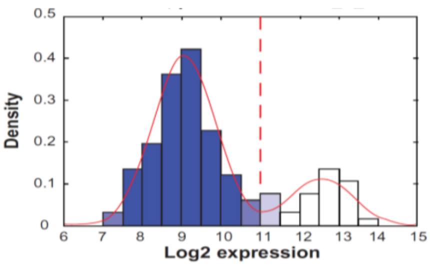

Figure 1: Histograms for ERBB2 gene. The gene has bimodal distribution with the dashed vertical line representing the classification threshold between the two modes [3].

Khalil et al have explained that cancer is a complex disease [4] because it has many subtypes . The existence of bimodal genes may be related to important subtypes of a disease. In medical science, bimodal genes can be the product of somatic mutations as the amplification of the receptor tyrosine kinase proto-oncogene "erbB2" during the development of cancer [5]. Another cause for the bimodality in cancers is germ cell mutations such as SNPs [6]. It has been noticed that the majority of cancer data demonstrate this kind of heterogeneous pattern [7-9]. Genetic translocations are commonly occurred in cancer cell which is a result of the rearrangement of parts between non-homologous chromosomes [10]. However, these mutations play main role in cancer cell progression or, more generally, diseases development. Furthermore, the genomic lesions may affect some samples but not all leading to the occurrence of bimodality. An example of recurrent fusion was observed by Tomlins and others in prostate cancer datasets where they found ERG and ETV1 genes over expressed in some of the samples in multiple datasets [9]. A study has showed that oncogene HER2 is over-expressed in 15–20% of breast tumors compared with normal breast tissues [11]. In addition the bimodality appears in biological systems as noticed by Mason and his group [12]. It is observed that the expression levels for some genes showed a distinct bimodal distribution in human skeletal muscle tissue. Also bimodal distribution were studied in blood glucose samples [13, 14]. The bimodality can occur in humongous tissue as reported in these references [12, 15].

This heterogeneity demonstrated that the fully understanding to both genotype and phenotypes is the critical key for drug design [8]. The researchers have made a great effort to study the complexity of cancer disease aiming to understand the molecular characteristics [16, 17]. Cancer patients with similar tumour characteristics are likely not to response for the same treatment [18]. In breast cancer, for example, variant responses were found to drug such as Tamoxifen and Herceptin giving evidence of the heterogeneity in pathological factors such as estrogen receptor (ER) and HER2 status [19]. Large number of patients gained from using Tamoxifen for hormone receptor-positive but the same drug failed in subgroup of patients who carry specific variants in the cytochrome gene P450 2D6 (CYP2D6) [20, 21]. Trastuzumab, as a first drug approved by FDA for this purpose, has been a beneficial therapy, either alone or in combination with chemotherapy, in about 25% of patients with positive HRE2 cancer patients [22-25]. This raised an issue of an accurate grouping of HRE2-positive patients [21]. Gefitinib (Iressa) has been approved by FDA, which suppress the ATP binding function of EGFR, and has been of partially remission regression for 10-30% of patients with non-small cell lung cancer [26-30]. It has been noticed, that genetic alterations are associated with drug response as proven in their study [31].

Due to the often observed heterogeneity in gene expression data,the conventional *t* test and correlation analysis may not be able to well detect partial differentiation. The kurtosis analysis [32], the likelihood ratio test [33] and the bimodality index [34] have been proposed to examine the bimodality among genes. PACK (Profile Analysis using Clustering and Kurtosis) [32] clusters samples first and then uses kurtosis to find relevant classifiers. It was reported that about 80%-20% bimodal genes were missed using PACK [34]. The likelihood ratio test (LRT) [33, 35] examines the likelihood of bimodal over unimodal [13, 14]. Ertel and Tozeren used the χ^2 test with six degree of freedoms. They set 0.001 as the significance level to predict bimodal genes. Bessarabova and colleagues developed a τ indicator for detecting bimodality [36]. They combined a statistical method based around *t* test like statistic for direct comparison of gene expression from different platforms to identify bimodal genes based on the relative difference average between each peak of gene expression value in breast cancer. The Bimodality Index [34] used a mixture of two homogeneous Gaussians to model bimodality and outweighed the high-expressed samples.

Have applied these algorithms to our data, we have found that they often show dissatisfied performance. Some often over-predict bimodal genes and some do not provide a statistical significance value for analysis. In this paper we present a novel algorithm further. The basic principle is to detect the maximum gaps between two clusters. This therefore avoids the parametric function to be used. We have evaluated this algorithm in comparison with several benchmark algorithms and demonstrate in this paper that this new algorithm provides another way to acquire insightful interpretation to bimodality among genes.

In the following sections, we discuss the implementation of hBI and evaluate its performance in comparison to some benchmark algorithms using real and simulated data.

2 Methods

Our algorithm is a revision of Bimodal Index - BI [34], which is defined as:

$$\mathrm{BI_i} = \sqrt{\pi_i(1-\pi_i)}\delta_i \qquad (1)$$

where π_i is the proportion of samples and δ_i is the distance between the two subgroups of the i^{th} genes. The use of this definition implies a homogeneous variance for two clusters of samples. A one-side *t* statistics of the i^{th} gene can be defined as

$$t_i = \frac{\mu_{H,i} - \mu_{L,i}}{\sqrt{\dfrac{\sigma^2_{L,i}}{n_{L,i}} + \dfrac{\sigma^2_{H,i}}{n_{H,i}}}} \qquad (2)$$

where $\sigma^2_{L,i}$ is the variance of lowly expressed samples, $\sigma^2_{H,i}$ is the variance of highly expressed samples, $n_{L,i}$ is the number of lowly expressed samples, $n_{H,i}$ is the number of highly expressed samples, $\mu_{L,i}$ is the mean of lowly expressed samples, and $\mu_{H,i}$ is the mean of highly expressed samples of the i^{th} gene. if $\sigma^2_{L,i} = \sigma^2_{H,i} = \sigma^2_i$, this one side *t* statistic becomes

$$t_i = \sqrt{\frac{n}{\sigma^2_i}}\delta_i\sqrt{\pi_{H,i}(1-\pi_{H,i})} \qquad (3)$$

where $\pi_{H,i}$ is the proportion of highly expressed samples of the i^{th} gene. If the sample size is fixed for all genes,

$$t_i \propto \sigma_i^{-1}\delta\sqrt{\pi_{H,i}(1-\pi_{H,i})} \qquad (4)$$

It can be seen that if homogeneous exists across subgroups and genes, BI is equivalent to one side *t* statistic. However this can hardly be true in real applications. We therefore revise BI employing heterogeneous variance. In the one side *t* statistic, we use percentile estimations to replace parametric estimation of means and variances shown below

$$t_i = \frac{q^{25}_{H,i} - q^{75}_{L,i}}{\sqrt{\dfrac{\sigma^2_{L,i}}{n_{L,i}} + \dfrac{\sigma^2_{H,i}}{n_{H,i}}}} \qquad (5)$$

Here q^{25}_H is the 25th percentile of highly expressed samples, q^{75}_L is the 75th percentile of lowly expressed samples and the variances are calculated using

$$\sigma = \frac{\mathrm{IQR}}{1.34896} \qquad (6)$$

We assume that the separation between lowly expressed samples and highly expressed samples occurs at one of the largest gaps between consecutive sorted samples. Therefore we introduce the gap between lowly expressed samples and highly expressed samples to enhance the bimodality test. Our heterogeneous bimodal index (hBI) is defined below

$$\mathrm{hBI_i} = \alpha(m_{H,i} - M_{L,i}) + (1-\alpha)t_i \qquad (7)$$

where $m_{H,i}$ is the minimum of highly expressed samples, $M_{L,i}$ is the maximum of lowly expressed samples of the i^{th} gene and $\alpha > 0$ is a trade-off between the gap effect and *t* statistic. In this paper, $\alpha = 0.75$.

BI uses an arbitrary threshold to make decision based of the indexes, we employ the sequential Monte Carlo approach [37] (Besag and Clifford 1996) to deliver significance analysis. The procedure of our algorithm is shown below

Step 1. BI calculation for each gene

1.1. to sort expressions

1.2. to calculate the distance between every consecutive expressions and record them as a gap list
1.3. to sort the gap list
1.4. to calculate the revised BI for the top ten gaps and record them in a bimodality list
1.5. to maximise the bimodality list

Step 2. Apply BC algorithm to obtain *p* values

To evaluate our algorithm in comparison with likelihood test, Kurtosis test and BI test, we calculate sensitivity (Sen), specificity (Spe), total accuracy (Auc) and use receiver operative characteristic (ROC) [38] analysis. The sensitivity is the ratio of correctly predicted bimodal genes. The specificity is the ratio of correctly predicted non-bimodal genes. The total accuracy is the ratio of corrected identified unimodal and bimodal genes. Specially, we calculate the area under ROC curve (AUC) for comparison.

3 Results and discussions

3.1 Simulated Data

For all five scenarios, 950 genes were designed as unimodal and 50 genes were designed as bimodal. Each gene has 40 replicates. Thirty replicates were designed of low expressions. Ten replicates were designed of high expressions. Each simulation was repeated for ten times.

Scenario 1- Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed samples of a bimodal gene were drawn from a normal distribution of mean ten and standard deviation one. Highly expressed samples of a bimodal gene were drawn from a normal distribution of mean 12 with variable standard deviation drawn from a uniform distribution between one and five. *Table 1* shows the comparison based on the mean values among ten simulations for four algorithms using specificity, sensitivity and AUC. It can be seen that hBI and Kurtosis have similar performance and hBI slightly outperforms Kurtosis analysis. Likelihood test shows the worst performance with the sensitivity as 0.06 although its specificity is 1.

Table 1: The averaged measurements for scenario 1

	LR	K	BI	hBI
Spe	1	0.983	0.975	0.992
Sen	0.062	0.858	0.532	0.84
Auc	0.995	0.964	0.852	0.992

Scenario 2- Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed samples of a bimodal gene were drawn from a normal distribution of mean ten and standard deviation one. Highly expressed replicates of a bimodal gene follow a uniform distribution in the interval between zero and five in addition to maximum of low expressions. The averaged measurements are shown in *Table 2*. In this scenario kurtosis has shown the worst accuracy (36%) while the other relatively similar and higher, 99.9%. Also the result has shown that the likelihood test has very low sensitivity while BI and hBI perform equally well.

Table 2: The averaged measurements for scenario 2

	LR	K	BI	hBI
Spe	1	0.986	0.997	0.9963
Sen	0.058	0	0.954	0.924
Auc	0.999	0.36	0.999	0.9985

Scenario 3 **-** Samples of unimodal genes were drawn from a uniform distribution in the interval between ten and 12. Lowly expressed replicates of a bimodal gene were drawn from the same low expression distribution as bimodal genes and highly expressed replicates of a bimodal gene were drawn from a normal distribution with two units added to the maximum of the low expressions. *Table 3* shows the summary of the simulations for this scenario. This scenario has shown that Kurtosis failed again to have a sensible accuracy (14%). hBI shows the highest AUC (0.999) similar to LR (0.996) and BI (0.993). hBI outweighs LR and BI in term of sensitivity, the sensitivities of BI and LR are 0.81 and 0.77, respectively while hBI's sensitivity is 0.94.

Table 3: The averaged measurements for scenario 3

	LR	K	BI	hBI
Spe	0.997	0.9519	0.9898	0.996
Sen	0.774	0	0.812	0.94
Auc	0.996	0.1413	0.9933	0.999

Scenario 4 **-** Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. Lowly expressed replicates of a bimodal gene were drawn from a mixture of a normal distribution of mean ten and a normal distribution of mean 12. The standard deviation of the former was designed as one and that of the latter was designed as three. Highly expressed replicates of a bimodal gene were drawn from the low expressions plus white noise with two units above the maximum low expression. ***Table 4*** shows the summary of ten simulations on random samples for this scenario. All perform very well in terms of AUC. This means there are some suitable statistical significance levels by which perfect separation between unimodal and bimodal genes can be found.

Table 4 The averaged measurements for scenario 4

	LR	K	BI	hBI
Spe	1	0.9861	0.9963	0.997
Sen	0.468	0	0.93	0.952
Auc	0.998	0.9572	0.9984	0.9988

Scenario 5 **-** Samples of unimodal genes were drawn from a normal distribution of mean ten and standard deviation one. We organised lowly expressed replicates of a bimodal gene as a mixture of three normal distributions with mean values as ten, 11 and 12 as well as standard deviation values as three, two and one. Highly expressed replicates of a bimodal gene were drawn in the same way as scenario 4. Based on ten random simulations for this scenario, we have observed that although LR and BI show reasonably good values of AUC, their sensitivities are not acceptable. This shows that these two algorithms have the same problem encountered in scenario 4 that their p values tend to be large, which leads to the difficulty of using command significance levels to make decision. Kurtosis analysis does not work well because its AUC value drops to 0.66 not very far away from 0.5, a random classification. In this scenario hBI perform the best in all measurements while BI has 69% sensitivity.

Table 5: The averaged measurements for scenario 5

	LR	K	BI	hBI
Spe	1	0.9844	0.9845	0.9873
Sen	0	0	0.698	0.766
Auc	0.9267	0.6676	0.9593	0.9867

3.2 Real data

GSE11121 dataset: The data set was downloaded from GEO (Gene Expression Omnibus). It contains 200 lymph node-negative breast cancer patients who were not treated by systemic therapy after surgery. The data was derivation study to find prognostic motifs [39]. Gene expression profiling of patients was done using the Affymetrix HG-U133A microarray platform compromising 22283 probs. The raw expression deposited at the NCBI GEO data repository under the accession number GSE11121. We have transformed the expression using base two logarithm before analysis. We used three significance levels (0.001, 0.01 and 0.05) to predict bimodal genes. *Table 6* shows the predicted bimodal genes using these three significance levels. The likelihood test predicted from 0.3% to 2.3% bimodal genes, BI predicted from 0.01% to 5% bimodal genes and hBI predicted bimodal genes from 0.01% to 5% as well. However Kurtosis analysis ends up with too many predictions up to 54.7%, which is unreasonable. Even for the significance level 0.001, it still predicts 36.3% bimodal genes, which is far more than a realistic level.

Table 6: Number of predicted bimodal genes for three significance levels for data set GDS11121

Significance levels	LHR	K	BI	hBI
0.001	72	8087	22	23
0.01	182	10065	227	221
0.05	523	12193	1112	1112

Fig 2 (a) shows the overlap analysis between four algorithms based on the significance level 0.001 values using VennDiagram [40]. We have found that hBI is most similar to BI. The overlap percentage between these two algorithms is 31.8%, i.e. 100*7/(7+14+1). The overlap degree between LHR and hBI is 20.2%. The overlap degree between LHR and BI is 5.6%. 91.3% of predicted bimodal genes of hBI are predicted by Kurtosis as well. This percentage drops to 69.6% between hBI and LHR as well as 30% between hBI and BI. Also the overlap percentage between BI and hBI is 27.7% for significance level 0.01 and the overlap degree is 34.3% between the hBI and LHR and 7.9% between BI and LHR - *Fig 2* (b). 90.9% of predicted bimodal genes of hBI is predicted by Kurtosis as well. This percentage drops to 46.6% between hBI and LHR as well as 24% between hBI and BI. For significance level 0.05, we found the overlap between hBI and BI is 36.2% and the overlap degree between hBI and LHR is 28.2% and 7.07% between BI and LHR - *Fig 2* (c). In addition, 83.3% (32.3%, 36.2%) of hBI's predictions are consistent with Kurtosis (LHR, BI).

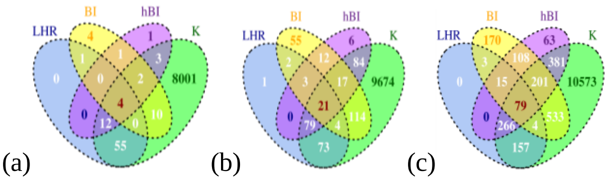

Fig 2: Venn diagram illustrates the overlapped between the methods for GSE11121 with the significance levels 0.001(a),0.01(b) and 0.05(c).

Fig 3 shows top five bimodal genes predicted based on the significance level 0.001, where (a-d) predicted by all and (e) was predicted by hBI only. It can be seen that they show different types of distributions. Both GOXA1 - *Fig 3* (a) - and GATA3 - *Fig 3* (b) show a pattern that the high expressions form a tight cluster. However the low expressions demonstrate a more flat distribution or form more small clusters. TDRD12 - *Fig 3* (c) - and GRIA2 - *Fig 3* (d) have tight clusters formed by low expressions and their high expressions display flat distributions. SH3GL3 shows a different pattern from other four. It is composed of two more tightly formed clusters, one small and one large. The gap between two clusters is large. The analysis of these patterns proves one important concept that the use of restrict assumption of data distribution may not be sufficient for accurate prediction of bimodal genes in real applications, where distribution can vary in many different formats.

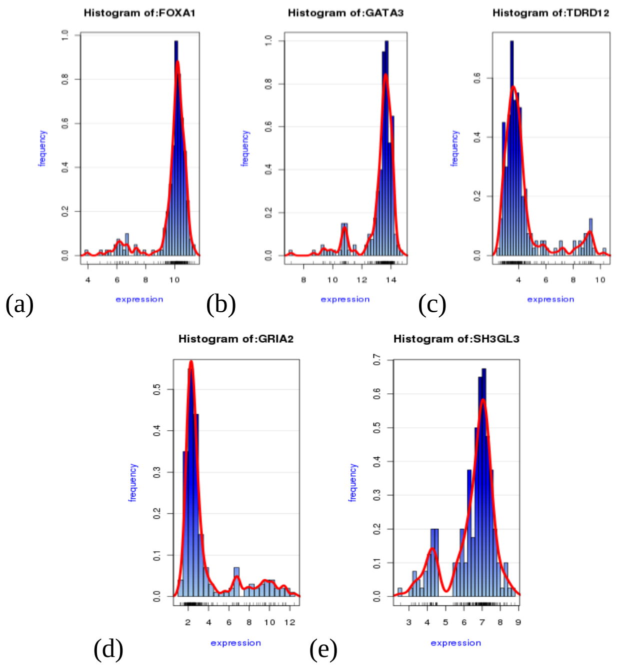

Fig 3: Density analysis of four bimodal genes; (a-d) predicted by all four algorithms at the significance level 0.001 and (e) only predicted by hBI at the same significance level. The horizontal axes represent $\log_2$ expressions and the vertical axes represent frequencies. All these genes show typical bimodal (or multi-modal) distributions.

Table 7 shows the *p* values of four algorithms for the genes uniquely predicted by hBI at the significance level 0.01. The data shows that for those bimodal genes predicted by hBI, their ranks of other algorithms are far behind. For instance, the Kurtosis rank of C6orf64 is 18643 and the Kurtosis rank of GULP1 is 22101. Fig 4 shows four of them, which have gene symbols. They are indeed bimodal genes. However other three algorithms failed to predict them. For instance, PSPH was ranked by hBI at the 66th position (p = 0.002). Likelihood, Kurtosis and BI ranked it at the 2990th (p = 0.2), 13507th (p = 0.1), and the 1293th (p = 0.05) respectively. This gene is highly expressed in African Americans comparing to European Americans colorectal cancer patients [41]. Also PSPH is expressed at higher level in responding patients versus non-responding group, which support its importance as therapeutic target for non-small-cell lung cancer [42]. RBBP5 was ranked at the 113th position (p = 0.005), but was ranked at the 540th position (p = 0.52), the 17000th position (p = 0.36), and the 974th position (p = 0.043) by likelihood, Kurtosis and BI tests. RBBP5 was found to be active in only 40% of Pancreatic ductal adenocarcinomas (PDAs) [43].

Table 7: *p* values of bimodal genes predicted **ONLY** by hBI at significance level 0.01 for GDS11121

symbol	LH	K	BI	hBI
PSPH	0.27(2990)	0.1(13507)	0.058(1293)	0.002(66)
unknown	0.065(652)	0.03(11438)	0.039(864)	0.004(97)
RBBP9	0.052(540)	0.36(17000)	0.043(974)	0.005(113)
unknown	0.31(3584)	0.02(11311)	0.012(265)	0.008(195)
C6orf64	0.07(713)	0.54(18643)	0.018(416)	0.008(199)
GULP1	0.19(1879)	0.97(22101)	0.026(582)	0.009(226)

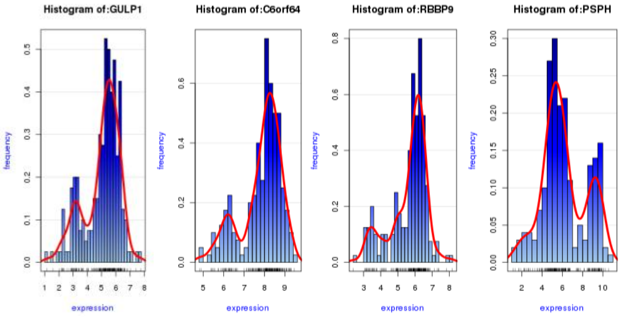

Fig 4. Density analysis of two bimodal genes only predicted by hBI at the significance level 0.01. The horizontal axes represent $\log_2$ expressions and vertical axes represent frequencies. All these genes show typical bimodal (or multi-modal) distributions.

4 Conclusion

We have proposed a novel bimodal gene prediction algorithm via relaxing the constraints of BimodalIndex algorithm. First, the constraint of cross-cluster homogeneous variance has been removed. It is unrealistic to assume that two clusters of a bimodal gene should have the same variance. The examination of various data sets has clearly shown that one of two clusters, either being of lowly expressed samples or of highly expressed samples is very likely to demonstrate a comparatively flat distribution while the other shows a tight cluster. Second, we deliberately removed the constraint of homogeneous variance across genes because this constraint is certainly confusing. An obviously evidence is that the variance of unimodal genes and bimodal genes will not show homogeneous variance. In addition to these two revisions, we have also emphasised the impact of gaps between consecutive expressions of sorted samples on bimodal formulation. This is because we have observed in real data sets that often lowly expressed samples demonstrate a tight cluster and highly expressed samples show; *i*) a comparatively large variance; and *ii*) distantly departing from the tight cluster of lowly expressed samples or vice versa. In this case the *t* statistic, although using percentiles to estimate mean values and standard deviations, is still not working well, i.e. the *t* statistic can be very likely to be small due to the large variance of the highly expressed samples. We therefore introduced a gap impact onto the prediction of bimodal genes. Doing so, we admit that we have introduced a hyper-parameter. In order to remove this hyper-parameter, our future work will employ the Bayesian learning framework to overcome this difficulty. Nevertheless, we have documented our simulations, which all show that our new algorithm is better than the benchmark algorithms in simulated data sets. In the application to real data sets, we show that our new algorithm is partially consistent with benchmark algorithms and does provide some new insights to the analysis of bimodal genes. Importantly, most of the predicted bimodal genes by our new algorithm do show typical bimodality. Particularly, not a small percentage of our unique predictions is unfortunately not favoured by benchmark algorithms. We therefore look forward to some even advanced approach, such as meta-analysis of prediction to deliver even robust predictions of bimodal genes. Finally, it worth to note that significance analysis is critical to real biological/medical application, we therefore have enhanced the BimodalIndex for using the Besag's sequential Monte Carlo approach to deliver significance analysis.

5 References

[1] J. L. DeRisi, Iyer, Vishwanath.R., Brown, Patrick O., "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science,* vol. 278, pp. 680-686, 1997.

[2] Y. Yang, Tashman, Adam., Lee, Jung., Yoon, Seungtai., Mao, Wenyang., Ahn, Kwangmi., Kim, Wonkuk., Mendell, Nancy., Gordon, Derek., Finch, Stephen., "Mixture modeling of microarray gene expression data," *BMC Proceedings,* vol. 1, p. S50, 2007.

[3] A. Ertel, "Bimodal Gene Expression and Biomarker Discovery," *Cancer Informatics,* vol. 9, pp. 11-14, 2010.

[4] I. G. Khalil, Hill, C., "Systems biology for cancer," *Current Opinion in Oncology,* vol. 17, pp. 44-48, 2005.

[5] J. G. Hengstler, Lange, Jost., Kett, Alexandra., Dornhöfer, Nadja., Meinert, Rolf., Arand, Michael., Knapstein, Paul G., Becker, Roger., Oesch, Franz., Tanner, Berno., "Contribution of c-erbB-2 and Topoisomerase IIα to Chemoresistance in Ovarian Cancer," *Cancer Research,* vol. 59, pp. 3206-3214, 1999.

[6] V. N. Kristensen, Edvardsen, Hege., Tsalenko, Anya., Nordgard, Silje H., Sorlie, Therese., Sharan, Roded., Vailaya, Aditya., Ben-Dor, Amir., Lonning, Per Eystein., Lien, Sigbjorn., Omholt, Stig., Syvanen, Ann-Christine., Yakhini, Zohar., Borresen-Dale, Anne-Lise., "Genetic variation in putative regulatory loci controlling gene expression in breast cancer," *Proceedings of the National Academy of Sciences,* vol. 103, pp. 7735-7740, 2006.

[7] W. F. Anderson, Matsuno, Rayna., "Breast Cancer Heterogeneity: A Mixture of At Least Two Main Types?," *Journal of the National Cancer Institute,* vol. 98, pp. 948-951, 2006.

[8] F. B. Bertucci, Daniel., "Reasons for breast cancer heterogeneity," *Journal of Biology,* vol. 7, p. 6, 2008.

[9] S. A. Tomlins, Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J.E., Shah, R.B., Pienta, K.J., Rubin, M.A., Chinnaiyan, A.M., "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science,* vol. 310, pp. 644-8, 2005.

[10] J. Hu, "Cancer outlier detection based on likelihood ratio test," *Bioinformatics,* vol. 24, pp. 2193-9, 2008.

[11] D. Slamon, Clark, GM., Wong, SG., Levin, WJ., Ullrich, A., McGuire, WL., "Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene," *Science,* vol. 235, pp. 177-182, 1987.

[12] C. Mason, Hanson, Robert., Ossowski, Vicky., Bian, Li., Baier, Leslie., Krakoff, Jonathan., Bogardus, Clifton., "Bimodal distribution of RNA expression levels in human skeletal muscle tissue," *BMC Genomics,* vol. 12, p. 98, 2011.

[13] T.-O. B. Lim, Rugayah. Morad, Zaki. Hamid, Maimunah A., "Bimodality in Blood Glucose Distribution: is it universal?," *Diabetes Care,* vol. 25, pp. 2212-2217, 2002.

[14] J. M. Fan, Susanne.J. Zhou, Yue. Barrett-Connor, Elizabeth., "Bimodality of 2-h Plasma Glucose Distributions in Whites," *Diabetes Care,* vol. 28, pp. 1451-1456, 2005.

[15] I. K. Dozmorov, Nicholas. Tang, Yuhong. Shields, Alan. Pathipvanich, Parima. Jarvis, James,N. Centola, Michael., "Hypervariable genes—experimental error or hidden dynamics," *Nucleic Acids Research,* vol. 32, p. e147, 2004.

[16] C. Blenkiron, Goldstein, Leonard., Thorne, Natalie., Spiteri, Inmaculada., Chin, Suet-Feung., Dunning, Mark., Barbosa-Morais, Nuno., Teschendorff, Andrew., Green, Andrew., Ellis, Ian., Tavare, Simon., Caldas, Carlos., Miska, Eric., "MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype," *Genome Biology,* vol. 8, p. R214, 2007.

[17] S. Chin, Teschendorff, Andrew., Marioni, John.,Wang, Yanzhong., Barbosa-Morais, Nuno., Thorne, Natalie., Costa, Jose., Pinder, Sarah., van de Wiel, Mark., Green, Andrew., Ellis, Ian., Porter, Peggy., Tavare, Simon., Brenton, James., Ylstra, Bauke., Caldas, Carlos., "High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer," *Genome Biology,* vol. 8, p. R215, 2007.

[18] M. Gort, Broekhuis, Manda., Otter, Renée., Klazinga, Niek., "Improvement of best practice in early breast cancer: actionable surgeon and hospital factors," *Breast Cancer Research and Treatment,* vol. 102, pp. 219-226, 2007.

[19] R. E. Ellsworth, Hooke,Jeffrey.A., Shriver,Craig.D., Ellsworth,Darrell.L. , "Genomic Heterogeneity of Breast Tumor Pathogenesis," *Clinical Medicine Insights: Oncology* vol. 3, pp. 77-85, 2009.

[20] L. D. Bradford, "CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants," *Pharmacogenomics,* vol. 3, pp. 229-243, 2002.

[21] R. E. D. Ellsworth, David.J.; Shriver, Craig.D.; Ellsworth, Darrell.L. , "Breast Cancer in the Personal Genomics Era," *Current Genomics,* vol. 11, pp. 146-161, 2010.

[22] D. B. A. Agus, Robert W.; Fox, William D.; Lewis, Gail D.; Higgins, Brian.; Pisacane, Paul I.; Lofgren,

Julie A.; Tindell, Charles; Evans, Douglas P.; Maiese, Krista; Scher, Howard I.; Sliwkowski, Mark X., "Targeting ligand-activated ErbB2 signaling inhibits breast and prostate tumor growth," *Cancer Cell,* vol. 2, pp. 127-137, 2002.

[23] M. Harris, "Monoclonal antibodies as therapeutic agents for cancer," *The Lancet Oncology,* vol. 5, pp. 292-302, 2004.

[24] R. E. Nahta, Francisco., "HER2 therapy: Molecular mechanisms of trastuzumab resistance," *Breast Cancer Research,* vol. 8, p. 215, 2006A.

[25] R. E. Nahta, Francisco.J., "Herceptin: mechanisms of action and resistance," *Cancer Letters,* vol. 232, pp. 123-138, 2006B.

[26] B. A. R. Chabner, Thomas G., "Chemotherapy and the war on cancer," *Nat Rev Cancer,* vol. 5, pp. 65-72, 2005.

[27] M. G. Muhsin, Joanne.; Kirkpatrick, Peter., "Gefitinib," *Nat Rev Cancer,* vol. 3, pp. 556-557, 2003.

[28] P. A. E. Jänne, Jeffrey A.; Johnson, Bruce E., "Epidermal Growth Factor Receptor Mutations in Non–Small-Cell Lung Cancer: Implications for Treatment and Tumor Biology," *Journal of Clinical Oncology,* vol. 23, pp. 3227-3234, 2005.

[29] M. G. Kris, Natale, Ronald B.,Herbst, Roy S., Lynch, Thomas J., Prager, Diane., Belani, Chandra P., Schiller, Joan H., Kelly, Karen., Spiridonidis, Harris., Sandler, Alan., Albain, Kathy S., Cella, David., Wolf, Michael.K., Averbuch, Steven.D., Ochs, Judith.J., Kay, Andrea.C., "Efficacy of Gefitinib, an Inhibitor of the Epidermal Growth Factor Receptor Tyrosine Kinase, in Symptomatic Patients With Non–Small Cell Lung Cancer," *JAMA: The Journal of the American Medical Association,* vol. 290, pp. 2149-2158, 2003.

[30] H.-C. C. Wu, De-Kuan.; Huang, Chia-Ting., "Targeted Therapy for Cancer," *J. Cancer Mol.,* vol. 2, pp. 57-66, 2006.

[31] J. G. J. Paez, Pasi A. Lee, Jeffrey C.; Tracy, Sean.; Greulich, Heidi.; Gabriel, Stacey.; Herman, Paula.; Kaye, Frederic J.; Lindeman, Neal.; Boggon, Titus J.; Naoki, Katsuhiko.; Sasaki, Hidefumi.; Fujii, Yoshitaka.; Eck, Michael J.; Sellers, William R.; Johnson, Bruce E.; Meyerson, Matthew., "EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy," *Science,* vol. 304, pp. 1497-1500, 2004.

[32] A. E. Teschendorff, Naderi, A., Barbosa-Morais, N.L., Caldas, C., "PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer," *Bioinformatics,* vol. 22, pp. 2269-75, 2006.

[33] A. Ertel, Tozeren, A., "Switch-like genes populate cell communication pathways and are enriched for extracellular proteins," *BMC Bioinformatics,* vol. 9, p. 3, 2008.

[34] J. Wang, Wen, S., Symmans, W.F., Pusztai, L., Coombes, K.R., "The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data," *Cancer Inform,* vol. 7, pp. 199-216, 2009.

[35] M. Gormley, Tozeren, A., "Expression profiles of switch-like genes accurately classify tissue and infectious disease phenotypes in model-based classification," *BMC Bioinformatics,* vol. 9, p. 486, 2008.

[36] M. Bessarabova, Kirillov, E., Shi, W., Bugrim, A., Nikolsky, Y., Nikolskaya, T., "Bimodal gene expression patterns in breast cancer," *BMC Genomics,* vol. 11, p. S8, 2010.

[37] J. BESAG and P. CLIFFORD, "Sequential Monte Carlo p-values," *Biometrika,* vol. 78, pp. 301-304, 1991.

[38] C. E. Metz, "Basic principles of ROC analysis. ," *Seminars in Nuclear Medicine,* vol. 8, pp. 283-288, 1978.

[39] M. Schmidt, Böhm, Daniel., von Törne, Christian., Steiner, Eric., Puhl, Alexander., Pilch, Henryk., Lehr, Hans-Anton., Hengstler, Jan G., Kölbl, Heinz.,Gehrmann, Mathias.,, "The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer," *Cancer Research,* vol. 68, pp. 5405-5413, 2008.

[40] H. Chen, Boutros, Paul, "VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R," *BMC Bioinformatics,* vol. 12, p. 35, 2011.

[41] B. Jovov, Araujo-Perez, Felix., Sigel, Carlie S., Stratford, Jeran K., McCoy, Amber N., Yeh, Jen Jen., Keku, Temitope., "Differential Gene Expression between African American and European American Colorectal Cancer Patients," *PLos ONE,* vol. 7, p. e30168, 2012.

[42] E.-H. Tan, Ramlau, R., Pluzanska, A., Kuo, H.-P., Reck, M., Milanowski, J., Au, J. S.-K., Felip, E., Yang, P.-C., Damyanov, D., Orlov, S., Akimov, M., Delmar, P.,Essioux, L., Hillenbach, C., Klughammer, B., McLoughlin, P. Baselga, J., "A multicentre phase II gene expression profiling study of putative relationships between tumour biomarkers and clinical response with erlotinib in non-small-cell lung cancer," *Annals of Oncology,* vol. 21, pp. 217-222, 2010.

[43] D. J. Shields, Niessen, Sherry., Murphy, Eric A., Mielgo, Ainhoa., Desgrosellier, Jay S., Lau, Steven K. M., Barnes, Leo A., Lesperance, Jacqueline., Bouvet, Michael., Tarin, David., Cravatt, Benjamin F., Cheresh, David A.,, "RBBP9: A tumor-associated serine hydrolase activity required for pancreatic neoplasia," *Proceedings of the National Academy of Sciences,* vol. 107, pp. 2189-2194, 2010.