Vol.2012-BIO-28 No.13
2012/3/28

Refined blood-borne miRNome of human diseases via PCA-based feature extraction

Y-h. Taguchi[†1] and Yoshiki Murakami[†2]

Disease biomarker using blood is clinically important, since blood is easy to obtain from patients, thus it requires relatively less stress. However, blood generally reflects not only targeted diseases but also whole body status of patients. Thus, it is important which contents of blood are considered. Recently, miRNAs in blood, blood-borne miRNome, turns out to be promising candidates for blood based biomarker for diseases. In this paper, we propose a new method based upon principal component analysis to identify better candidates for miRNAs as blood based biomarker using miRNA expression profiles of patients.

Our method based upon principal components analysis provides us better blood-borne miRNome to discriminate diseases from healthy controls. They are hsa-miR-425,hsa-miR-15b,hsa-miR-185, hsa-miR-92a, hsa-miR-140-3p, hsa-miR-320a, hsa-miR-486-5p, hsa-miR-16, hsa-miR-191, hsa-miR-106b, hsa-miR-19b, and hsa-miR-30d and are previously extensively reported to be cancer/disease related miRNAs. We have found that these common miRNAs are expressive or suppressive significantly in most of diseases/cancers, but in diseases/cancers specific combinatory manner. It enables us to discriminate cancers/diseases from healthy control well.

1. Introduction

Specific and sensitive non-invasive biomarkers for the detection of human epithelial malignancies are urgently required to reduce the worldwide morbidity and mortality caused by cancer. Recently, circulating microRNAs are realized to be a new candidate for clinical biomarker[1)–5)]. MicroRNAs are post-transcriptional regulators that are involved in many physiological and pathophysiological conditions. A recent study compared the expression profiles of hundreds of blood-borne microRNAs across a variety of nonmalignant and malignant diseases to identify disease-specific expression patterns. The resulting microRNA expression data could be used to discriminate disease samples with a high level of accuracy, demonstrating the potential for using microRNA signatures for the blood-based diagnosis of disease.

Recently, Ref. 6) proposed blood-borne miRNome of human diseases. They have shown wide range of diseases/cancers is discriminated from healthy control by only miRNA expression using extensive bioinformatics research.

In spite of that, selecting biomarker based upon feature extraction techniques is still an issue[7)]. Although Ref. 6) successfully discriminates diseases/cancers from healthy control using only 10 miRNAs expression very accurately, they did not present which 10 miRNAs are selected. This is possibly because of the problem pointed out by Ref. 7), stability. Although Ref. 6) employed ten-fold cross validation, the selection of 10 miRNAs possibly fluctuated from one training set to another training set. This prevented them from presenting 10 miRNAs for biomarker to discriminate each cancer/disease from healthy control.

In this paper, we propose a new feature selection technique to select miRNAs as biomarker based upon principal component analysis (PCA).

2. Materials and Methods

2.1 Feature extraction based upon PCA

Suppose we have miRNA profiles $x_{ij}, (i = 1, \ldots, N, j = 1, \ldots, M)$, each of which corresponds to ith miRNA in jth sample. Samples are classified into L clinical sets, $G_l, (l = 1, \ldots, L)$. Then we have applied PCA to the set of $\{x_{ij}\}$ in two ways;

(1) Method 1 (miRNA based): Substitute $K_s(< M)$ principal component (PC) score x_{ik} to x_{ij}. In this case, PCA is applied to a matrix $\{x_{ij}\}$.

(2) Method 2 (sample based): Substitute $K_m(< N)$ principal component (PC) score x_{kj} to x_{ij}. In this case, PCA is applied to a transverse matrix $\{x_{ji}\}$.

The PCA based feature extraction is as follows;

(1) Step one : Choose a pair of clinical sets, l and l'.

(2) Step two : Compute x_{ik} with method 1 PCA from $\{x_{ij} \mid j \in G_l \cup G_{l'}\}$.

†1 Department of Physics,Chuo University
†2 Center for Genomic Medicine, Kyoto University Graduate School of Medicine

(3) Step three : Compute distance r_i,

$$r_i \equiv \sqrt{\sum_{k=1}^{K_s^0} x_{ik}^2},$$

where $K_s^0 (< K_s)$ is the number of components to be used for feature selection.

(4) Step four : Select miRNAs i' with top $N_1 (< N)$ r_is.

N_1 miRNAs are a set of selected features to distinguish clinical sets l and l'. Thoughout this paper, K_s^0 is constantly taken to be 2, if not explicitly denoted. PCA is computed by prcomp function in R[8] base package.

One should notice that PCA based feature extraction do not make use of classification information at all. This method is classification free method and is very unique because of this point (see Discussion section).

2.2 The PCA based linear discriminant analysis

The PCA based linear discriminant analysis (LDA) is as follows;

(1) Step one : Choose a pair of clinical sets, l and l'.

(2) Step two : If necessary, apply a feature extraction and reduce number of miRNAs used for LDA.

(3) Step three : Compute $x_{kj}, (k = 1, \ldots, K_m)$ using method 2 PCA.

(4) Step four : Divide sampled into training set and test set.

(5) Step five : Apply LDA to training set.

(6) Step six : Validate the performance of LDA using test set.

(7) Step seven : Repeat steps from four to six many times.

(8) Step eight : Compute performance with averaged values.

One should notice that division between training and test sets are done **AFTER** computation of PCA (and feature extraction if necessary). Thus, x_{kj} include the information of test sets, too. Feature extraction, if applied, is also before division, thus is sampling free. One may think that it is a fake since we do not know classification of test set. However, even if we do not have preknowledge about classifications, we can compute PCA, since we do not need classification information to compute x_{kj}. This will be discussed at discussion section, too. LDA is computed by lda function in R[8] base package.

2.3 Simulation data set

In order to evaluate the performances of proposed method, we have generated artificial data set as follows. Suppose we have two pseudo clinical sets G_1 and G_2. Samples $j \leq (>) \frac{M}{2}$ belong to $G_1(G_2)$. miRNAs $i \leq N_1$ are supposed to have distinct values between G_1 and G_2 and others are not as follows;

$$x_{ij} = \begin{cases} N(\mu_{i1}, D\sigma_{i1}) & j \leq \frac{M}{2}, \quad i \leq N_1 \\ N(\mu_{i2}, D\sigma_{i2}) & j > \frac{M}{2}, \quad i \leq N_1 \\ N(\mu_i, \sigma_i) & i > N_1 \end{cases},$$

where $N(\mu, \sigma)$ is the normal distribution having mean μ and standard deviation (SD) of σ. $\mu_{i1}, \mu_{i2}, \mu_i, \sigma_{i1}, \sigma_{i2}, \sigma_i$ are taken from uniform distribution $\in [0, 1]$. $D (= 0.1, 0.2, \ldots, 2.0)$ is the parameter to represent how easy G_1 and G_2 are discriminated. Larger (smaller) D means harder (easier) to discriminate between two pseudo clinical sets.

2.4 miRNA expression and normalization

The miRNA expression used in this study is taken from Gene Expression Omnibus (GEO) having accession number of GSE31568, which was used in Ref. 6)'s study. We have downloaded GSE31568_raw and normalized miRNA expression within each sample so as to have zero mean and unit SD.

2.5 The amount of contribution by each miRNA to discriminations

Suppose we have x_{kj} by PCA analysis after PCA based feature extraction is applied. Then

$$x_{kj} = \sum_{i=1}^{K_m} a_{ik} x_{ij}$$

If we apply LDA to discriminate one of cancers/diseases from healthy control using x_{kj}, we get discriminant function LD_j as

$$LD_j = \sum_{k=1}^{PC} b_k x_{kj} = \sum_{k=1}^{PC} b_k \sum_{i=1}^{K_m} a_{ik} x_{ij} = \sum_{i=1}^{K_m} \left(\sum_{k=1}^{PC} b_k a_{ik} \right) x_{ij},$$

for jth sample, while PC is the number of PCs used for discrimination. Typically, positive (negative) LD_j means sample j belongs to cancers/diseases (healthy control) sample. Then amount of contribution C_i of miRNA i to the discriminant function is

$$C_i = \sum_{k=1}^{PC} b_k a_{ik}.$$

3. Results

3.1 Simulation Results

N and M are taken to be 100 and 200 respectively and $N_1 = 10$. First, we have generated x_{ij} one hundred times for each of $D, (= 0.1, 0.2, \ldots, 2.0)$. Any performances are evaluated by averaged values over one hundred ensembles. Next, we try to discriminate between G_1 and G_2 by LDA. Cross validations are done by LOOCV (leave-one-out cross validation, Fig. 1A). Averaged accuracy ranges from 0.70 to 1.0. Thus, G_1 and G_2 can be discriminated well within this range of D values. An interesting thing is that LDA which used $x_{ij}, (i < N_1)$ outperformed LDA with all. Since $x_{ij} (i < N_1)$ are the expressions where G_1's means and SDs differ from G_2's, it turns out that selecting informative components only is useful to improve performance. Thus, if feature extraction can select informative components correctly, performance will be improved.

In Fig. 1B, we have shown how well each feature extraction method can select correct miRNAs (i.e., those with $i \leq N_1$) when D varies. We have applied both PCA based and P-value based feature extractions for each of D while generating 100 ensembles and have chosen N_1 miRNAs. When D is small enough, P-values, which is computed by t test, based feature extraction[6] can correctly select almost all informative components. In contrast to this, PCA based feature selection cannot select as many as informative components for $D < 1.0$. However, one should notice that it still can select more than half of them correctly. When D exceeds 1.0, PCA based feature extraction method starts to outperform P-value based feature extraction. Here we also compute the number of miRNAs having false discovery rate FDR[9] < 0.05, it deceases as D increases. Since the number is upper limit of correctly selected miRNAs by P-value based method and that by PCA based feature extraction exceeds it when $D > 1.0$, PCA based method is definitely better than P-value based method for $D > 1$. The number of correctly selected miRNAs by P-value based method continuously decreases as D increase while that by PCA based method increases. Since we do not know which situation has happened because we can never know which miRNAs should be selected, we conclude that it is safer to employ PCA based method than P-value based method.

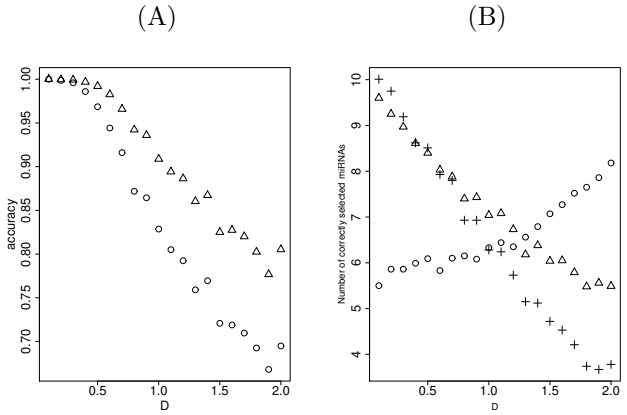

Fig. 1 (A) Accuracy of LDA between G_1 and G_2 using simulation data as a function of D. ∘ is the result obtained using all miRNAs and △ is the result obtained using N_1 miRNAs ($i \leq N_1$). (B) Number of correctly selected miRNAs by P-value based (△) and PCA based feature extraction (∘). + shows total number of miRNAs with FDR < 0.05.

3.2 Biomarker identification for diseases and cancers

We have applied PCA based feature extraction to biomarker identification for diseases and cancers[6]. As will be discussed in Discussion section, since our PCA based feature extraction is free from sampling, we can strictly define top 10 miRNAs which is distinct between a pair of one of clinical traits and a healthy control (Table 1). We can also make use of these selected miRNAs for discrimination between diseases/cancers from healthy control as it is. Table 2 is the performance of PCA based LDA between diseases/cancers and healthy controls using only these selected 10 miRNAs. It is competitive to or slightly better than Ref. 6)'s results (see p14 of their Supplementary materials).

4. Discussion

In contrast to Ref. 6), we have successfully listed 10 miRNAs for biomarker. The reason why they could not do this is possibly because P-value based feature extraction is deeply dependent upon divisions between training and test sets. Since they have done 100 trials of division, it is unlikely for them to have definite set of 10 miRNAs (see below).

Table 1 miRNAs selected as biomarker to distinguish diseases/cancers from healthy control by PCA based feature extraction. (1)lung cancer, (2)other pancreatic tumors and diseases, (3) pankreatitis, (4) ovarian cancer, (5) copd, (6) pancreatic cancer ductal, (7) tumor of stomach, (8) sarco- idosis, (9) prostate cancer, (10)acure myocard infarction, (11)perio- dontitis, (12) multiple sclerosis, (13) mela- noma, and (14)wilms tumor. + (-) indicates it is expressive in cancers/diseases (healthy control), i.e., $C_i > (<)0$. $*$ means it is not selected within top ten most significant miRNAs which contribute to discriminations. A-C: miRNAs belong to common cluster. Coincidence within clusters A and C are underlined.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
A	425	+	+	+	-	+	+	-	+	+	-	-	-	-	-
B	15b	-	-	-	+	+	+	-	-	-	+	-	-	-	+
	185	-	-	-	-	-	-	-	+	-	-	-	-	+	-
C	92a	+	+	-	+	-	+	+	+	-	+	-	+	+	+
	140-3p	+	+	+	+	+	-	+	+	+	-	+	+	+	+
	320a	+	-	-	-	+	-	-	+	+	+	+	+	+	+
	486-5p	-	+	+	-	-	+	-	+	-	+	-	-	-	-
B	16	-	+	+	+	+	+	+	-	+	-	+	+	-	-
A	191	+	+	+	-	+	+	+	-	+	-	-	*	+	-
	106b	*	+	+	-	+	*	-	*	-	-	-	-	*	-
C	19b	+	*	*	*	*	+	*	*	*	*	*	-	+	*
	30d	*	*	*	*	*	*	*	+	*	*	*	*	*	*

Table 2 Performance of PCA based LDA for discrimination between diseases/cancers and healthy control. + (-) indicated that the performance is better(worse) than Ref. 6). PC is the number of PCs used for PCA based LDA.LOOCV is applied. See Ref. 6)'s Table at p14 of Supplementary materials.

diseases/cancers	PC	Accuracy	Specitivity	Sencitivity	Precision
lung cancer	5	0.784(+)	0.875(+)	0.750(+)	0.632
other pancreatic tumors & diseases	7	0.814	0.900(+)	0.875(+)	0.724
pankreatitis	8	0.833(+)	0.948(+)	0.921(+)	0.700
ovarian cancer	6	0.800	0.965(+)	0.867(+)	0.464
copd	2	0.713(-)	0.922(+)	0.833(+)	0.465
pancreatic cancer ductal	2	0.765(-)	0.852(+)	0.800(+)	0.667
tumor of stomach	9	0.855(+)	0.968(+)	0.846	0.524
sarcoidosis	10	0.835(-)	0.918(+)	0.889(-)	0.741
prostate cancer	5	0.806(+)	0.933(+)	0.826(+)	0.576
acure myocard infarction	7	0.789(+)	0.514(-)	0.757(-)	0.964
periodontitis	10	0.807(+)	0.934(+)	0.778(-)	0.519
multiple sclerosis	10	0.892(+)	0.984(+)	0.957(+)	0.710
melanoma	10	0.867(-)	0.938(+)	0.886(-)	0.756
wilms tumor	7	0.867	0.969	0.600	0.273

Instead of the list of 10 miRNAs used for discriminations, they have listed miRNAs that were deregulated in at least six diseases (Ref. 6)'s Supplementary

Table 3 miRNAs in Table 1 whose curated up/downregulations in any cancers are found in Ref. 10). Any miRNAs listed in Ref. 10)'s additional file 1. CNS : Central Nervous System.

miRNA	Cancer Type	Expression	Mean fold change
hsa-miR-425	CNS	Down regulated	13.6 fold reduction
hsa-miR-15b	Colon	Up regulated	˜1.5 fold increase
hsa-miR-185	Bladder(Urothelial)	Up regulated	1.30 fold increase
hsa-miR-185	Kidney	Up regulated	1.42 fold increase
hsa-miR-92-2	Pancreas	Up regulated	
hsa-miR-92-2	Prostate	Up regulated	
hsa-miR-140	CNS	Down regulated	2.7 fold reduction
hsa-miR-140	Colon	Down regulated	11.4 fold reduction
hsa-miR-140	Hematologic	Down regulated	3.5 fold reduction
hsa-miR-140	Lung	Down regulated	
hsa-miR-140	Ovary	Down regulated	3.51 fold reduction
hsa-miR-16-1	Uterus / Endometrial cancer	Up regulated	At least 2 fold increase
hsa-miR-16-2	B-Cell-CLL	Down regulated, Deleted	
hsa-miR-16a	B-Cell-CLL	Down regulated	
hsa-miR-191	Breast	Up regulated	
hsa-miR-191	CNS	Down regulated	4.4 fold reduction
hsa-miR-191	Colon	Up regulated	1.4 fold increase
hsa-miR-191	Lung	Up regulated	
hsa-miR-106b	Lung	Up regulated	12-fold increase in small lung cancer cell line SKLC-2.
hsa-miR-30d	CNS	Down regulated	3.2 fold reduction

Table 2). Surprisingly, overlap between them and our Table 1 is very little. The overlap between Table 1 and their 24 miRNAs that were significantly deregulated in > 50 % of all diseases is only hsa-miR-320a. Considering whole Supplementary Table 2 results in the addition of only one miRNA, hsa-miR-16. Even if we take into account their Figure 1 where upregulated miRNAs are considered together, no other miRNAs are found to be selected in both their paper and present study.

Recently Ref. 11) tried similar research by next generation sequencing. They have renewed a list of significant miRNAs in supplementary information, but again there are only two overlaps, miR-425 (for tumor of stomach and wilms tumor) and miR-140-3p (for melanoma, ovarian cancer and peridontitis).

In order to validate our selections independent of their research, we have checked if there are previous reports to support our findings that these miRNAs are deeply related to cancers/diseases. Then it turns out that miRNAs in Table 1 have huge number of previous published reports to support the relation-

ship with cancers/diseases (not shown here). Although all of the previous reports are not always coincident wth each other, miR-15b, miR-185, miR-140-3p, miR-320a, miR-486-5p, miR-16, and miR-30d turn out to work generally as tumor suppressors and miR-425, miR-92a, miR-191, miR-106b, and miR-19b are primary oncogenic. In order to confirm if our judge is valid, we have shown in Table 3 the curated up/downregulation of some miRNAs in several cancers[10]. First of all, since not all miRNAs have curated up/downregulation records, the fact that most of miRNAs in Table 1, excluding three miRNAs (miR-320, miR-486 and miR-191), are listed supports that our findings agree with previous knowledges. Their up/downregulation patterns are basically coincident with what we have denoted in the above, since tumor suppressor (oncogene) should be suppressive (expressive) in cancers. Some miRNAs among them have a little bit complicated functionalities. For example, miR-185 is frequently upregulted in cancers (see Table 3) while its expression sometimes suppresses cell proliferations. Another example of miRNAs with not-straghtforward feature is miR-15b. It is not always suppressive in tumors. At least it is reported in Table 3 to be upregulated in colon cancer. In pite of that, it sometimes inhibits tumor function. This not-so-easy-to-understand situation can be seen in expression profiles, too. Even if we see heatmap (not shown here), one can see no specific expression of miRNAs are associated with cancers/diseases. We need more sophisticated views more than observing individual miRNA expression one by one.

Anyway, if we also consider the fact that our list is common for the most of comparisons between healthy control and cancers/diseases, we believe that we can conclude that our list of miRNAs as biomarkers to distinguish between diseases/cancers and healthy controls is trustable. It is because such a coincidence hardly occurs by simple accidental agreements. There are too many miRNAs for it to occur accidentally. Our list is plausible even if it does drastically differ from Supplementary Table 2 by Ref. 6). Anyway, there are no theoretical/biological reasons that a set of 10 representative miRNAs to discriminate between cancers/diseases and healthy control must be unique.

In order to understand more deeply how each miRNA cooperatively discriminates between cancers/diseases and healthy control, we have listed contributes to discrimination by each of miRNAs (Table 1). Since LDA is a linear method, it allows us to do this easily (see Materials and Methods).

Interestingly, miRNAs which belong to the same cluster often share the combinations of positive/negative contributions. For example, as marked "C" in the left most column of Table 1 and underlined, there are remarkable coincidence between miR-92a and miR-19b. Three (lung cancer, pancreatic cancer ductal, and melanoma) out of four cancers/diseases where miR-19b's contributions are listed share the outcomes, although it is not significant ($P = 0.3125$). Similarly, miR-425 and miR-191 (marked "A" and underlined in Table 1) have the same positive/negative contributions for 10 cancers/diseases ($P = 0.046$) out of 13 cancers/diseases where miR-191 has non-zero contributions (three exceptions are, tumor of stomach, sarcoidosis, and melanoma). However, since this does not stand between miR-15b and miR-16 (marked "B" but not underlined in Table 1 because of small number of coincidences), this is again not-so-straightforward as expected.

Some miRNAs are coincident with their known functions. For example, miR-486-5p, which is known to be tumor suppressive miRNA (see above). It is more expressive in normal control (not shown here). On the other hand, miR-92a is more expressive in cancers/diseases side, which does not disagree with the previously known belief; miR-17-92 cluster is oncogene.

However, some other miRNAs show controversial features to previous knowledges. For example, miR-106b and miR-425 are believed to belong to oncogenic miRNAs, but is expressive mainly in normal control side (not shown here). These apparent discrepancies are possibly because miRNAs are not measured in tissues but in blood. If we check PhenomiR data base[12], we can find many cases that expression in blood differs from that in tissues. For example, miR-140 is reported to be downregulated in lung cancer (database ID 132 and 134) but is over expressive in lung cancer serum (database ID 503). miR-92a-1 is reported to be downregulated in lung cancer (database ID 530 and 543) while it is over expressive in lung cancer serum (database ID 503). These findings in blood is in agreement with the present study that miR-140 and miR-92a are shown to be expressive in lung cancer blood (Table 1). Similarly, miR-92a is highly expressed in hepatocellular carcinoma (HCC), but miR-92a in the plasmas from HCC patients is decreased compared with that from the healthy donors[13]. How expression in blood differs

from that in tissue is the next issue when using miRNAs in blood as biomarker.

In conclusion, although we cannot fully understand their features, when miRNAs in blood is used for biomarker to discriminate cancers/diseases from healthy control, at least, it is coincident with the previous proposal in tissue miRNAs[10]; significant features are not always expression of cancers/diseases specific miRNAs, but also expression of common miRNAs in cancers/diseases specific manner. More investigation along this line is waited.

5. Conclusion

In this paper, we have proposed a new feature extraction method based upon PCA for biomarker decision from miRNAs in blood. For simulation data, our method outperforms the conventional methods to detect informative components form the mixture of informative components and noise. When our method is applied to miRNA expression of diseases/cancers and normal control, we have found 10 common miRNAs independent of diseases/cancers considered. PCA based LDA using these 10 miRNAs can discriminate cancers/diseases from healthy control mostly competitively or slightly better than discrimination using 10 miRNAs selected by P-value based feature selection. It has been shown, for the first time, that most distinctive feature of cancers/diseases is not the expression of the specific miRNAs but the expression of mostly common miRNAs in the cancers/diseases specific manner.

Acknowledgement

This work was supported by KAKENHI (23300357).

References

1) Zen, K. and Zhang, C.Y.: Circulating MicroRNAs: a novel class of biomarkers to diagnose and monitor human cancers, *Med Res Rev* (2010).
2) Yu, D.C., Li, Q.G., Ding, X.W. and Ding, Y.T.: Circulating MicroRNAs: Potential Biomarkers for Cancer, *Int J Mol Sci*, Vol.12, pp.2055–2063 (2011).
3) Scholer, N., Langer, C. and Kuchenbauer, F.: Circulating microRNAs as biomarkers - True Blood?, *Genome Med*, Vol.3, p.72 (2011).
4) Brase, J.C., Wuttig, D., Kuner, R. and Sultmann, H.: Serum microRNAs as non-invasive biomarkers for cancer, *Mol. Cancer*, Vol.9, p.306 (2010).
5) Pritchard, C.C., Kroh, E., Wood, B., Arroyo, J.D., Dougherty, K.J., Miyaji, M.M., Tait, J.F. and Tewari, M.: Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies, *Cancer Prev Res (Phila)* (2011).
6) Keller, A., Leidinger, P., Bauer, A., Elsharawy, A., Haas, J., Backes, C., Wendschlag, A., Giese, N., Tjaden, C., Ott, K., Werner, J., Hackert, T., Ruprecht, K., Huwer, H., Huebers, J., Jacobs, G., Rosenstiel, P., Dommisch, H., Schaefer, A., Muller-Quernheim, J., Wullich, B., Keck, B., Graf, N., Reichrath, J., Vogel, B., Nebel, A., Jager, S.U., Staehler, P., Amarantos, I., Boisguerin, V., Staehler, C., Beier, M., Scheffler, M., Buchler, M.W., Wischhusen, J., Haeusler, S.F., Dietl, J., Hofmann, S., Lenhof, H.P., Schreiber, S., Katus, H.A., Rottbauer, W., Meder, B., Hoheisel, J.D., Franke, A. and Meese, E.: Toward the blood-borne miRNome of human diseases, *Nat. Methods*, Vol.8, pp.841–843 (2011).
7) Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P. and Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, *Bioinformatics*, Vol.26, pp.392–398 (2010).
8) R Development Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2010). ISBN 3-900051-07-0.
9) Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Royal Stat. Soc. B*, Vol.57, pp.289–300 (1995).
10) Bandyopadhyay, S., Mitra, R., Maulik, U. and Zhang, M.Q.: Development of the human cancer microRNA network, *Silence*, Vol.1, p.6 (2010).
11) Keller, A., Backes, C., Leidinger, P., Kefer, N., Boisguerin, V., Barbacioru, C., Vogel, B., Matzas, M., Huwer, H., Katus, H.A., Stahler, C., Meder, B. and Meese, E.: Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients, *Mol. BioSyst.*, Vol.7, pp.3187–3199 (online), DOI:10.1039/C1MB05353A (2011).
12) Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C. and Theis, F.J.: PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes, *Genome Biol.*, Vol.11, p.R6 (2010).
13) Shigoka, M., Tsuchida, A., Matsudo, T., Nagakawa, Y., Saito, H., Suzuki, Y., Aoki, T., Murakami, Y., Toyoda, H., Kumada, T., Bartenschlager, R., Kato, N., Ikeda, M., Takashina, T., Tanaka, M., Suzuki, R., Oikawa, K., Takanashi, M. and Kuroda, M.: Deregulation of miR-92a expression is implicated in hepatocellular carcinoma development, *Pathol. Int.*, Vol.60, pp.351–357 (2010).