Vol.2014-BIO-40 No.2
2014/12/18

Comparison of Hepatocellular Carcinoma miRNA Expression Profiling as Evaluated by Next Generation Sequencing and Microarray

Yoshiki Murakami[2,a)] Toshihito Tanahashi[3,4,b)] Rika Okada[4,c)] Hidenori Toyoda[5,d)]
Takashi Kumada[5,e)] Masaru Enomoto[2,f)] Akihiro Tamori[2,g)] Norifumi Kawada[2,h)]
Y-h. Taguchi[6,1,i)] Takeshi Azuma[2,j)]

Abstract: Although microarray has been an important tool that can perform extensive gene expression analyses, next generation sequencing (NGS) has recently arisen as an alternative methodology that can measure gene expression. In this paper, we have compared microarray and NGS quantitatively using microRNA measurements in hepatocellular carcinoma (HCC) and found that these two are coincident with each other. NGS also turned out to be used for biomarker between HCC and normal tissue if the recently proposed principal component analysis based unsupervised feature extraction was applied.

1. Introduction

Microarray has been a useful tool for the simultaneous measurements of gene expression in mRNA expression level. However, next generation sequencing (NGS) has become more and more important because NGS can be potentially more quantitative than microarray. On the other hand, because of long history of microarray, presently stored gene expression data mainly consist of microarray that contributed to several biological new findings. Thus, in order to relate biological knowledge between microarray and NGS, it is important to compare their measurements using same samples. Recently [1], considering microRNA (miRNA) expression in hepatocellular carcinoma (HCC), we have tested how coincident with each other these two methodology can be. Then we have found [1] that these two can be quantitatively coincident with each other even considering differential expression. In addition to this, we have demonstrated that miRNA measured by NGS can classify between HCC and normal controls if the recently proposed principal component analysis based unsupervised feature extraction was applied.

1 Corresponding Author
2 Department of Hepatology, Osaka City University Graduate School of Medicine, Osaka, Japan
3 Department of Medical Pharmaceutics, Kobe Pharmaceutical University, Kobe, Japan
4 Division of Gastroenterology, Department of Internal Medicine, Kobe University Graduate School of Medicine, Kobe, Japan
5 Department of Gastroenterology, Ogaki Municipal Hospital, Ogaki, Japan
6 Department of Physics, Chuo University, Kasuga, Bunkyo-ku, Tokyo 112–8851, Japan
a) m2079633@med.osaka-cu.ac.jp
b) tana@kobepharma-u.ac.jp
c) rokada@med.kobe-u.ac.jp
d) tkumada@he.mirai.ne.jp
e) takashi.kumada@gmail.com
f) enomoto-m@med.osaka-cu.ac.jp
g) atamori@med.osaka-cu.ac.jp
h) kawadanori@med.osaka-cu.ac.jp
i) tag@granular.com
j) azumat@med.kobe-u.ac.jp

2. Materials and Methods

2.1 Microarray and NGS data

Microarray data was uploaded to Gene Expression Omnibus (GEO) using GEO ID GSE31164. NGS data measured by MiSeq was uploaded to DNA Data Bank of Japan Sequence Read Archive using accession number DRA001067. They include 14 HCC tumor samples and matched 6 normal control samples. More detailed information about materials (e.g., patient information and sample retrievals), please refer to original paper [1].

2.2 Extraction of miRNA expression from NGS data

Fastq file are treated by miRDeep2 [2] software and expression of individual miRNA was extracted.

2.3 Data Normalization

In order to demonstrate how simple treatment can achieve good enough performance, we simply applied normalization such that each sample of NGS and microarray data has zero mean and variance of one. More detailed pre-processing of NGS data and how to retrieve miRNA expression from microarray, see original paper [1].

2.4 Linear discriminant analysis using principal component analysis

Discrimination between HCCs and normal controls for miRNA expression derived from NGS data set was performed using linear discriminant analysis (LDA) using principal component analysis (PCA). At first, PCA was applied to 11 selected miRNAs (see next subsection) and LDA was applied using obtained PCs. PCA was applied to all 20 samples. Then, LDA was used with leave one out cross validation. Thus, discrimination was performed in semi-supervised way.

2.5 Selection of miRNAs using PCA based unsupervised feature extraction

Recently proposed PCA based unsupervised feature extraction (FE) [1], [3], [4], [5], [6], [7], [8], [9], [10] was applied to miRNA expression derived from NGS data. Samples are embedded into two dimensional space using PCA and 11 outliers are from the origin are extracted.

3. Results

3.1 Comparison between microarray and NGS

In order to investigate how coincident with NGS microarray is, several comparisons were performed. At first, the reproducibility of NGS was tested. In order that, same samples are sequenced multiple times. Fig. 1 shows the comparison between multiple sequencing of the same sample. Dependent upon samples considered, sequencing was performed twice or three times. It is clear that reproducibility was fairly well independent of considered samples. Thus, we can conclude that our procedure, MiSeq and miRDeep2, could successfully reproduce miRNA expression of HCC.

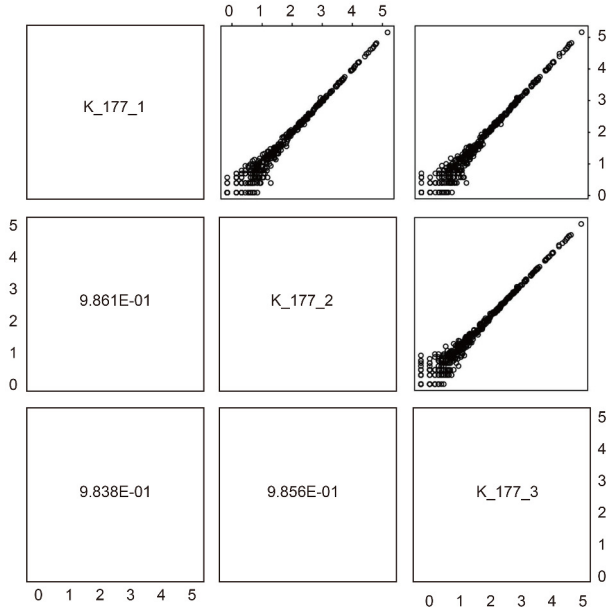

Fig. 1 Comparison of logarithmic miRNA expression between three technical replicates of HCC sample (patient ID K_177) measured by NGS analysis.

Since we have confirmed the reproducibility of NGS measurements, we also checked the reproducibility of differential miRNA expression (Fig. 2). Usually, expression is analysed in differential forms, e.g., logarithmic differential expression between tumor and normal tissues. Thus, reproducibility of differential expression is more important than that of raw expression.

Although Fig. 2 exhibits the differential expression between two distinct HCC samples that are expected to have relatively smaller differential expression between tumor and normal tissue, the reproducibility is fairly well. Among 36 pairwise comparisons, the smallest correlation coefficients were still larger than 0.8. Thus, it is expected for NGS to be coincident with microarray even considering differential expression.

Since we have successfully confirmed that MiSeq with miRDeep2 analysis has suitable reproducibility for the further analysis, we have compared miRNA expression between NGS and microarray. Fig. 3 represents the comparison of miRNA expression between NGS and microarray. Since the correlation coefficient 0.6059 is much lower than that obtained between three technical replicates (Fig. 1), it is still good enough. One drawback of NGS is insufficient number of reads that results in vertical array of points along vertical axis. This suggests that expression of rare miRNAs was truncated. Since MiSeq's ability for sequencing was updated since this study was performed, we can expect that this truncation will vanish if we can use more updated version of MiSeq.

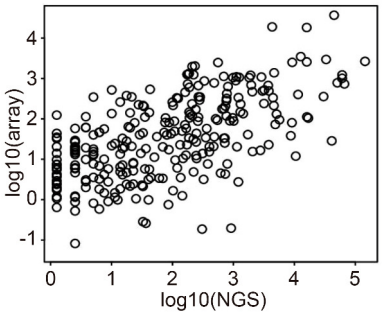

Fig. 3 Comparison of logarithmic miRNA expression between NGS and microarray, for the first replicate in Fig. 1. Person's correlation coefficient is 0.6059.

Since we have found that our methodology, MiSeq+miRDeep2, successfully reproduced miRNA expression measured by microarray, we further try to see if differential logarithmic expression measured by microarray can be reproduced by NGS.

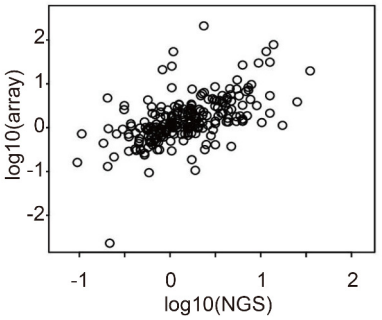

Fig. 4 Comparison of logarithmic differential miRNA expression between NGS and microarray, for the first pair (the top left) of replicates in Fig. 2. Person's correlation coefficient is 0.5555.

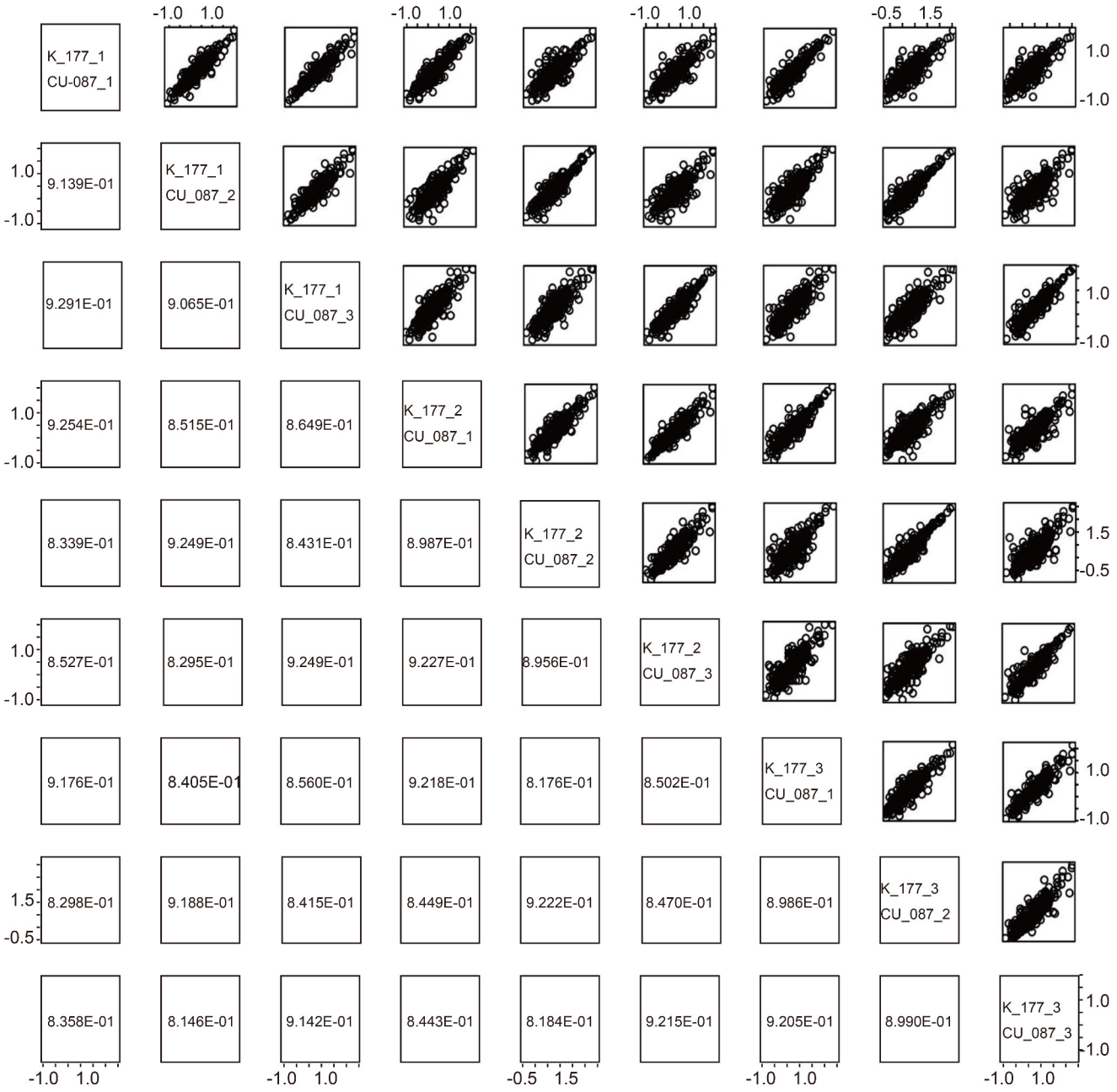

Fig. 2 Comparison of logarithmic differential miRNA expression between three technical replicates of HCC sample (patient IDs K_177 and CU_087) measured by NGS analysis.

Fig. 4 shows the comparison of the logarithmic differential miRNA expression between NGS and microarray, for the first pair (the top left) of replicates in Fig. 2. Although the correlation coefficients further deceased compared with that of logarithmic miRNA expression, the value 0.5555 is still large enough to conclude that our methodology at least qualitatively reproduced miRNA expression measured by microarray.

In spite of the simpleness of our treatment, NGC can reproduce miRNA expression measure by microarray relatively well.

3.2 Discrimination between HCC and normal controls

Although we have demonstrated the usefulness of our methodology, MiSeq + miRDeep2, worked pretty well and could reproduce miRNA expression measured by microarray, it is more useful if we can also show the biological significance of NGS measurements. Although there are many possibilities for this, we have selected to show the usefulness of miRNAs as biomarker. As previously demonstrated [5], [7], [9], we have applied PCA based unsupervised FE in order to select miRNAs that can discrminate HCC and healthy controls. Then, we have selected 11 miRNA (miR-10a-5p, miR-122-5p, miR-146b-5p, miR-148a-3p, miR-192-5p, miR-22-3p, miR-26a-5p, and miR-27b-3p, miR-10b-5p, miR-143-3p, and miR-21-5p).

Fig. 5 shows the boxplots of selected miRNAs. Table 1 shows the results of discrimination. HCC and normal control was successfully discriminated using optimal number (6) of the first PCs

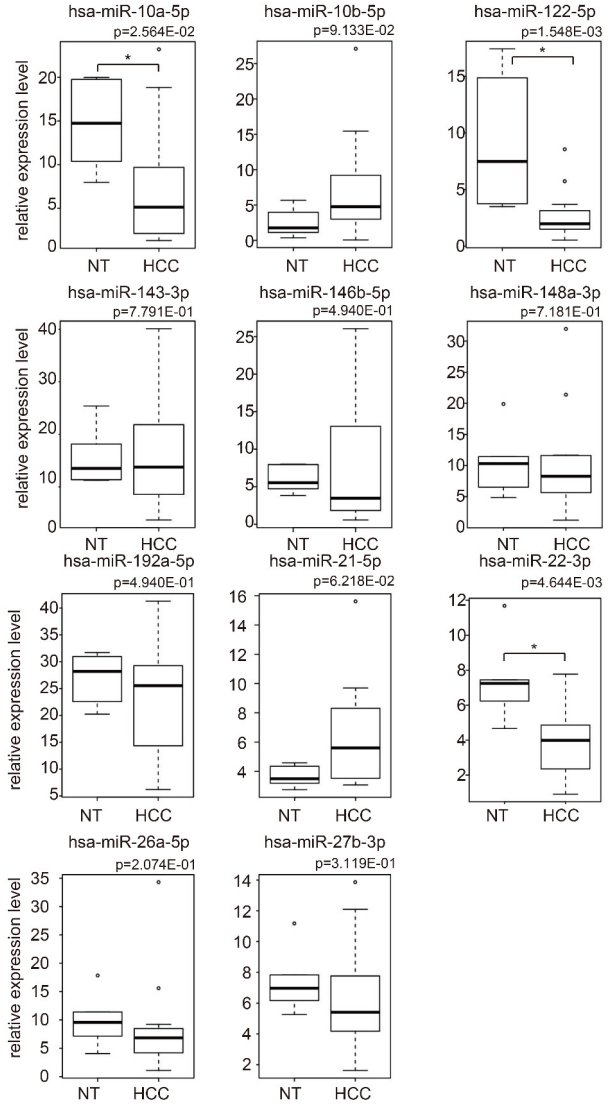

Fig. 5 Boxplot of miRNAs selected by PCA based unsupervised FE for the discrimination between HCC and normal controls. *P*-values were computed using two-sided Wilcoxon Rank Sum test. Asterisk indicates a significant difference of $p < 0.05$.

since accuracy was 0.90. *P*-values computed by Fisher's exact test is less than 7.22×10^{-4} and AUC is 0.92. Interestingly, in spite of the good performances, expression of selected 11 miRNAs are not always distinct between HCC and normal controls. Excluding two exceptions (miR-122-5p and 22-3p), no miRNAs are significantly distinct between HCC and normal controls. At a glance, this looks like a discrepancy. However, it it not a case, since LDA tried to discriminate between HCC and normal controls not using individual miRNA expression but using linear combination of miRNA expression. Thus, individual miRNA expression does not always have to be significantly distinct between HCC and normal controls. The reason why miRNAs used for discrimination were sought among those exhibiting significant difference between HCC and normal controls was simple because it is difficult to identify set of miRNAs that cannot discriminate between HCC and normal controls individually but can discriminate between them in a group. PCA based unsupervised FE has ability to easily identify miRNAs that can discriminate between HCC and normal controls without significant difference as individual miRNA [5], [7], [9].

Table 1 Discrimination between HCC and normal controls.

		Predicted	
		HCC	Normal
True	HCC	12	2
	Normal	0	6

3.3 Discovery of novel miRNAs

In addition to these above reported, we have also found a few novel miRNAs [1]. The criteria to identify novel miRNAs were 1) among the novel miRNAs identified by miRDeep2, those with a 80% probability of being a true positive, and 2) the miRNA was reproducibly detected in more than three samples. Based on the criteria, we have identified hsa-mir-9985, hsa-mir-1843, hsa-mir-548bc, and hsa-mir-9986 that will be included in the next release of miRBase [11]. The successful identification of novel miRNAs demonstrated the ability of miRDeep2 to identify unknown miRNAs from NGS data.

4. Conclusion

We have shown in this study that miRNA expression profiles obtained from NGS analysis are reproducible and are concordant with that obtained by the standard microarray procedure. Moreover, we have demonstrated that NGS can identify novel miRNAs that are otherwise undetectable by microarray analysis. HCC was distinguished from non-tumorous tissue with high diagnostic accuracy, supporting the clinical application of NGS- based miRNA expression profiling.

References

[1] Murakami, Y., Tanahashi, T., Okada, R., Toyoda, H., Kumada, T., Enomoto, M., Tamori, A., Kawada, N., Taguchi, Y. H. and Azuma, T.: Comparison of Hepatocellular Carcinoma miRNA Expression Profiling as Evaluated by Next Generation Sequencing and Microarray, *PLoS ONE*, Vol. 9, No. 9, p. e106314 (2014).

[2] Friedlander, M. R., Mackowiak, S. D., Li, N., Chen, W. and Rajewsky, N.: miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, *Nucleic Acids Res.*, Vol. 40, No. 1, pp. 37–52 (2012).

[3] Umeyama, H., Iwadate, M. and Taguchi, Y.-h.: TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer, *BMC Genomics*, p. in press (2014).

[4] Taguchi, Y.-h.: Integrative Analysis of Gene Expression and Promoter Methylation during Reprogramming of a Non-Small-Cell Lung Cancer Cell Line Using Principal Component Analysis-Based Unsupervised Feature Extraction, *Intelligent Computing in Bioinformatics*, LNCS, Vol. 8590, Springer, Heidelberg, pp. 445–455 (2014).

[5] Taguchi, Y. H. and Murakami, Y.: Universal disease biomarker: can a fixed set of blood microRNAs diagnose multiple diseases?, *BMC Res Notes*, Vol. 7, p. 581 (2014).

[6] Taguchi, Y.-h. and Okamoto, A.: Principal Component Analysis for Bacterial Proteomic Analysis, *Pattern Recognition in Bioinformatics*, LNCS, Vol. 7632, Springer Berlin Heidelberg, pp. 141–152 (2012).

[7] Murakami, Y., Toyoda, H., Tanahashi, T., Tanaka, J., Kumada, T., Yoshioka, Y., Kosaka, N., Ochiya, T. and Taguchi, Y. H.: Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease, *PLoS ONE*, Vol. 7, No. 10, p. e48366 (2012).

[8] Ishida, S., Umeyama, H., Iwadate, M. and Taguchi, Y. H.: Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery, *Protein Pept. Lett.*, Vol. 21, No. 8, pp. 828–39 (2014).

[9] Taguchi, Y. H. and Murakami, Y.: Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers, *PLoS ONE*, Vol. 8, No. 6, p. e66714 (2013).

[10] Kinoshita, R., Iwadate, M., Umeyama, H. and Taguchi, Y. H.: Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets, *BMC Syst Biol*, Vol. 8 Suppl 1, p. S4 (2014).

[11] Kozomara, A. and Griffiths-Jones, S.: miRBase: annotating high confidence microRNAs using deep sequencing data, *Nucleic Acids Research*, Vol. 42, No. D1, pp. D68–D73 (online), DOI: 10.1093/nar/gkt1181 (2014).