Vol.2016-BIO-48 No.1
2016/12/9

miRNA-mRNA相互作用同定を用いた腎芽腫関連遺伝子の推定

田口　善弘[1,a)]

概要：腎芽腫は、小児の腎腫瘍の一つであり、転移のない場合の 5 年生存率は 90 %以上であるが、発生機構などは依然、不明である。また、乳がん・卵巣がんの場合の BRCA1 の様な腫瘍形成に本質的な変異遺伝子も見つかっていない。本研究計画では、非コード RNA の一種である miRNA による mRNA の発現制御機構の異常が腎芽腫の発生原因であるという仮説に基づき、mRNA と miRNA の発現プロファイルの統合解析を行って、腎芽腫において特に mRNA-miRNA の関係（ 相関）が大きく変化しているペアを同定した。同定された mRNA のうち、複数個の miRNA の標的になっていると思われる mRNA の大部分は予後解析（ 生存曲線）に於いて腎臓がんに対して有意な P 値を持っていることが判明した。このことから、これらの遺伝子の発現異常が腎芽腫の発生に関係しているものと結論づけた。

Inference of Wilms timor causing genes using miRNA-mRNA interaction ideitification

Y-h. Taguchi[1,a)]

1. はじめに

腎芽腫は小児腎臓がんの一種であり、早期に発見されれば5 年生存率は９０％のがんである。しかし、一方で変異があると腎芽腫が発生するような原因遺伝子は特定されていない [1]。現在はプロモーターメチル化などのエピジェネティックな効果も含めて腎芽腫の生成要因が精力的に探索されている [2]。本研究では非コード RNA の一種である microRNA(miRNA) による mRNA の転写後発現制御の異常が腎芽腫の原因ではないかという仮説に基づき、mRNA,miRNA 発現プロファイルの統合解析を行い、原因となる mRNA を特定した。それらの多くは腎臓がんの生存曲線と有意に関係していることがわかった。このことからがん関連遺伝子の同定に miRNA-mRNA 相互作用を用いることは非常に有効であると考えられる。

2. 手法

全体の流れは図 **1** に描いた。

2.1 mRNA/miRNA プロファイル

mRNA/miRNA プロファイルは GEO [5] の GSE66405/GSE57370 から R [6] 内の GEOquery パッケージ [7] に実装されている getGEO 関数を使って R に読みこんだ。各プロファイルはサンプルごとに平均0，分散1 に規格化してから使用した。

2.2 主成分分析を用いた教師なし学習による変数選択

miRNA-mRNA 相互作用の同定の前に、腫瘍／正常臓器間で発現差がある miRNA/mRNA のスクリーニングを行った。その手法としては当該目的に有効であることが示されている [8] 主成分分析を用いた教師なし学習による変数選択法 [9–24] を用いた。本手法についての詳細は原著論

[1] 中央大学理工学部物理学科
Department of Physics, Chuo University, Tokyo 112–8551, Japan
[†1] 現在，情報処理大学
Presently with Johoshori Uniersity
[a)] tag@granular.com

この研究は IEEE BIBE2016 の Proceedings の一部 [3, 4] として既発表である

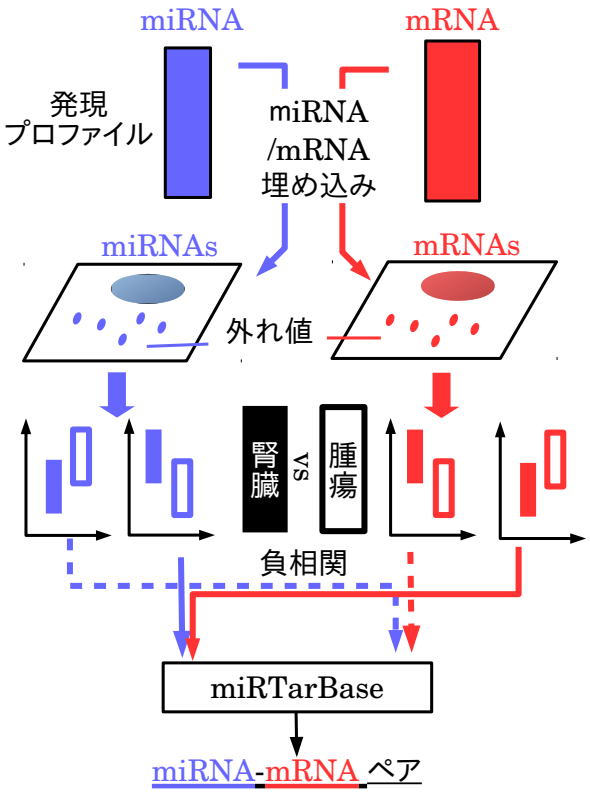

図 1　全体の流れ。まず、主成分分析を用いた教師なし学習による変数選択を用いて、mRNA/miRNA 別々に選択を行う。具体的には mRNA と miRNA を低次元に埋め込み、外れ値になっている mRNA と miRNA を選択。さらに選択された mRNA と miRNA のうちから、腫瘍／正常臓器間で有意に発現差があるものを選ぶ。これらのうち、miRTarBase に登録されている負相関のペアを miRNA-mRNA ペアとして同定する。

Fig. 1　Workflow of this study. miRNA/mRNA expression profiles were separately embedded into low dimensional space by PCA (feature embedding). After identifying PCs used for FE, outlier miRNAs/mRNAs are selected. miRNAs/mRNAs exhibiting significant differential expression between tumor and normal kidney were further selected among those selected as outliers, and pairs associated with reciprocal expression were compared with those listed in miRTarBase.

文 [3, 4] を見ていただくとして概要を述べる。まず、第一に主成分分析を用いて遺伝子を低次元空間に埋め込む。この結果、サンプルに主成分得点が、遺伝子に主成分負荷量が割り当てられる通常のサンプル埋め込みとは逆に、遺伝子に主成分得点が、サンプルに主成分負荷量が付与される。通常のサンプル埋め込みとの違いは、サンプル埋め込みの場合は共分散行列、または、相関係数行列が対角化されるために遺伝子ごとに発現プロファイルの平均がゼロになるのに対して、遺伝子埋め込みの場合はグラム行列が対角化されるためにサンプルごとの平均プロファイルがゼロになる点である。また、一般に、サンプル平均、遺伝子ごと平均のぞれぞれをゼロにした場合の効果は自明ではなく、この2 つの主成分分析は一般には異った結果を与えることになり、どの様に異なるかを事前に計算することもできない点は注意を要する。また、サンプル平均と遺伝子平均を同時にゼロにすることはできないので、どちらか一方を選択することになる。主成分分析を用いた教師なし学習による変数選択において、サンプル平均をゼロにする遺伝子埋め込みの方を選択する理由は、一般に遺伝子発現プロファイルの計測では絶対値の計算は不可能であり、なんらかの正規化が必要なためである。遺伝子平均をゼロにするサンプル埋め込みの場合は、前処理として遺伝子発現プロファイルがサンプルごとに何らかの仮定に基づいてあらかじめ正規化されていることが多い。しかし、そこに任意性が入ってしまったのでは、せっかくの教師なし学習の意味がない。そこで、通常、主成分分析を用いた教師なし学習による変数選択ではサンプル平均をゼロにする方が選択される。

次に、遺伝子選択に用いる主成分を主成分負荷量を参考に選択する。今の場合、主成分負荷量はサンプルに紐付けられているために、主成分負荷量を観測することでどの主成分に生物学的に意味がある特徴が表れているかを知ることができる。本研究の場合には、腫瘍／正常臓器間で発現差がある主成分負荷量を選んだ。

最後に、選択された主成分負荷量に対応する主成分得点に多重ガウス分布を仮定し、χ 二乗分布を適用して P 値を計算し、外れ値となっている遺伝子を選択する (多重比較補正された P 値が 0.01 以下を外れ値とする)。なぜなら、このような遺伝子は注目する主成分に対して、偶然では説明できない大きな寄与を持っている、つまり、周期性に特異的に寄与している遺伝子であるとみなすことができるからである。主成分得点がガウス分布に従うという仮定の是非には議論があるものと思われるが、この様な仮定は確率主成分分析と呼ばれる、主成分分析の統計学習的フレームでの再解釈でも採用されており [25]、それほど現実から離れたものではないと思われる。

2.3　腫瘍／正常臓器間に有意に発現差がある mRNA/miRNA の絞り込み

§2.2 でスクリーニングされた mRNA/miRNA の中から更に t 検定で腫瘍／正常臓器間に差がある miRNA/mRNA を絞り込む。得られた P 値を BH 基準 [26] で多重比較補正して $P < 0.05$ のもののみを選択する。

2.4　miRTarBase との比較

miRTarBase [27] は実験的に確認された miRNA-mRNA 相互作用だけを収集した信頼度の高い miRNA-mRNA 相互作用データベースである。§2.3 でスクリーニングされた mRNA/miRNA のうち、miRTarBase に登録されているペアに該当するものを選択し (但し、mRNA と miRNA は逆

表 1　腫瘍／正常臓器間の判別分析。行が予測で列が正解。mRNA は第一主成分負荷量、miRNA は第一から第七主成分負荷量を用いた場合がベスト

Table 1　Discrimination between normal kidneys and Wilms tumors. Row: prediction, column: true classes. The first PC loading for mRNA ($L = 1$) and the first seven PC loadings for miRNA ($L = 7$) were employed for the discrimination.

	mRNA		miRNA	
	腎臓	腫瘍	腎臓	腫瘍
腎臓	4	1	4	1
腫瘍	0	27	0	61

相のもの、つまり、mRNA が腫瘍＞正常臓器なら miRNA は腫瘍＜正常臓器、逆に mRNA が腫瘍＜正常臓器なら miRNA は腫瘍＞正常臓器であるもの)、これを同定された miRNA-mRNA 相互作用とする。

2.5　腫瘍／正常臓器間の判別分析

§2.2 でスクリーニングされた mRNA/miRNA を使って判別分析を行う。まず、選択された mRNA/miRNA だけを用いて主成分分析を再度行って、主成分負荷量を計算する。計算された主成分負荷量(サンプルに付与) を用いて、腫瘍／正常臓器間の線形判別分析を行う。使用する主成分負荷量の数を変え、もっともよい判別の時を報告する。

3.　結果

3.1　主成分分析を用いた教師なし学習による変数選択の **miRNA/mRNA** プロファイル解析への応用

§2.2 で記述された通りの方法を §2.1 にある miRNA/mRNA プロファイルに適用した結果、５５種の miRNA と１１１４個の mRNA をコードしたプローブが選択されることがわかった。

3.2　判別分析

§2.5 にある方法で、§3.1 で選択された miRNA/mRNA のみを用いて腫瘍／正常臓器間の判別分析を行った結果が、表 **1** である。判別は非常によく、主成分分析を用いた変数選択は腫瘍／正常臓器間で発現に差がある miRNA/mRNA をきちっと選べていることが確認できる。

3.3　負相関を伴った **miRNA-mRNA** ペアの同定

図 **2** は図 1 に描かれた手順で最終的に選ばれた miRNA-mRNA ペアで構成されたネットワークである。特に mRNA が腫瘍＞正常臓器、miRNA が腫瘍＜正常臓器、の発現を持っている場合(mRNA ががん遺伝子で、miRNA が腫瘍抑制因子であり、がん遺伝子を標的とする miRNA の発現が低下したためにがんになったと解釈できる) は全体がつながった大きなネットワークを構成するようにペアが選ばれており、妥当な結果であることが伺われる。

4.　議論

4.1　ネットワークのコア **mRNA** は生存分析に有意の影響を持つ

図 2 に描かれたネットワークの生物学的な妥当性を検証するため、図 2 のハブ mRNA を OncoLnc [28] にアップロードした。OncoLnc は既存のデータから各種のがんの生存分析に個々の遺伝子が寄与しているかどうかを記録したデータベースである。

表 **2** はその結果である。残念ながら腎芽腫は含まれていないがほとんどの遺伝子が2 種類の腎臓がん(太字表記) で有意の生存分析への寄与を持っている。このことは単純な mRNA の発現差だけからがんに関係した遺伝子を同定するよりも、miRNA の標的となっているかどうかを考慮することでより的確ながん遺伝子の同定ができると言う事実を示すと共に、腎芽腫の発生原因が miRNA の発現異常に起因するという当初の仮定を支持する結果であるとみなすことができるだろう。

5.　まとめ

本研究計画では主成分分析を用いた教師なし学習による変数選択でスクリーニングした miRNA/mRNA の中から miRNA-mRNA ペアを選ぶことで、より生物学的に意味があるがん関連遺伝子を同定できることを示した。図 2 に示されたネットワークは今後の腎芽腫の発生機構の研究に役立ち、また、表 2 にある遺伝子はバイオマーカー／治療標的として有用であることが期待される。

原著論文 [3, 4] ではさらに多くの生物学的な考察を行っているので興味ある場合には参照されたい。

参考文献

[1] Chu, A., Heck, J. E., Ribeiro, K. B., Brennan, P., Boffetta, P., Buffler, P. and Hung, R. J.: Wilms' tumour: a systematic review of risk factors and meta-analysis, *Paediatr Perinat Epidemiol*, Vol. 24, No. 5, pp. 449–469 (2010).

[2] Tian, F., Yourek, G., Shi, X. and Yang, Y.: The development of Wilms tumor: from WT1 and microRNA to animal models, *Biochim. Biophys. Acta*, Vol. 1846, No. 1, pp. 180–187 (2014).

[3] Taguchi, Y. H.: microRNA-mRNA interaction identification in Wilms tumor using principal component analysis based unsupervised feature extraction, *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering* (Ng, K. L., ed.), Los Alamitos, California, IEEE, IEEE Computer Society, pp. 71–78 (2016).

[4] Taguchi, Y. H.: microRNA-mRNA interaction identification in Wilms tumor using principalcomponent analysis based unsupervised feature extraction, *bioRxiv*, (online), DOI: 10.1101/059295 (2016).

[5] NCBI: Gene Expression Omnibus, NCBI (online), available from ⟨https://www.ncbi.nlm.nih.gov/geo/⟩ (ac-

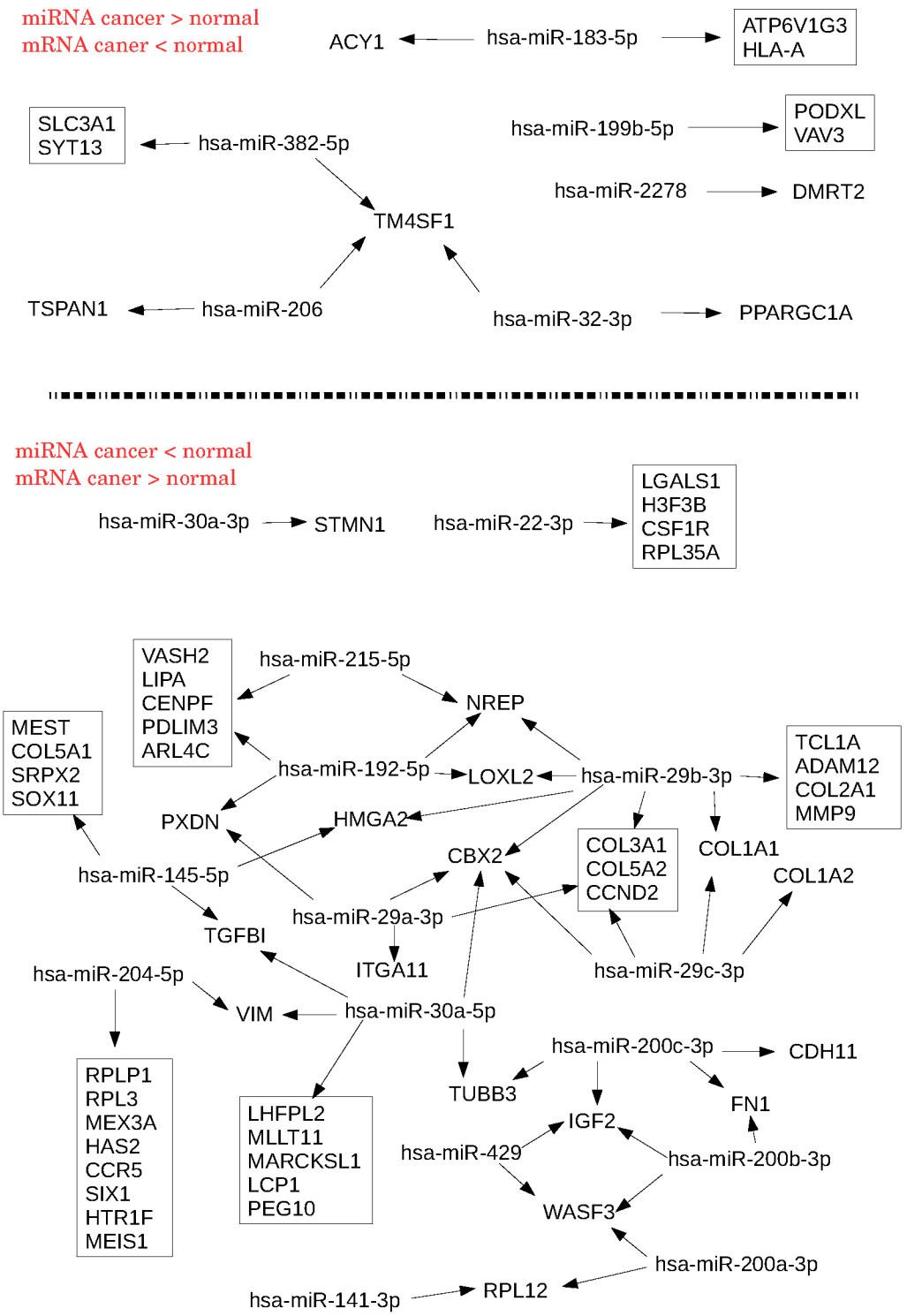

図 2　同定された miRNA-mRNA ペアから構成されるネットワーク
Fig. 2　miRNA-mRNA network composed of identified miRNA-mRNA pairs.

cessed 2016-11-08).
[6] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2015).
[7] Davis, S. and Meltzer, P.: GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor, *Bioinformatics*, Vol. 14, pp. 1846–1847 (2007).
[8] Taguchi, Y. H.: Identification of More Feasible MicroRNA-mRNA Interactions within Multiple Cancers Using Principal Component Analysis Based Unsupervised Feature Extraction, *Int J Mol Sci*, Vol. 17, No. 5,

表 2 OncoLnc による、各種腫瘍での生存確率への有意の関係。FDR 補正された P 値が 0.05 以下の mRNA。2 種の腎臓がんは太字で表記。LGG は腎臓がんと同時に発生することが多いので斜体で表記した

Table 2 Significant relationships to survival probabilities in various cancers provided by OncoLnc. Those associated with corrected FDR < 0.05. Two renal cancers are in bold. LGG was in italic in order to emphasize the association with renal cancers.

遺伝子	腫瘍	Cox	P 値	FDR 補正 P 値
4 種の miRNA に標的とされている mRNA				
CBX2	**KIRP**	0.908	2.90e-07	3.22e-05
	LIHC	0.501	1.70e-06	8.43e-04
	LGG	0.295	4.40e-03	1.22e-02
	KIRC	0.199	1.40e-02	4.29e-02
3 種の miRNA に標的とされている mRNA				
IGF2	**KIRP**	0.436	6.40e-03	3.67e-02
CCND1	**KIRC**	-0.248	3.60e-03	1.35e-02
COL3A1	**KIRP**	0.825	3.00e-06	1.71e-04
COL5A2	*LGG*	0.518	5.90e-07	8.14e-06
	KIRP	0.920	1.30e-07	1.63e-05
WASF3	*LGG*	-0.380	7.10e-05	3.84e-04
	SARC	0.394	7.70e-04	2.96e-02
NREp			—	
2 種の miRNA に標的とされている mRNA				
PXDN	**KIRC**	0.217	7.70e-03	2.55e-02
	CESC	0.615	4.70e-05	3.93e-02
HMGA2	**KIRC**	0.283	2.20e-04	1.57e-03
	PAAD	0.520	7.70e-06	4.20e-03
	KIRP	0.520	6.70e-04	7.81e-03
	SARC	0.373	3.10e-04	2.00e-02
	LUAD	0.234	2.50e-03	4.05e-02
LOXL2	*LGG*	0.342	6.50e-04	2.43e-03
	LUAD	0.298	6.80e-05	7.66e-03
	CESC	0.628	7.90e-06	2.46e-02
	KIRC	0.205	1.00e-02	3.11e-02
	KIRP	0.419	7.30e-03	4.03e-02
COL1A1	**KIRP**	0.881	4.90e-07	4.77e-05
	KIRC	0.252	1.80e-03	8.14e-03
	LGG	0.216	2.30e-02	4.88e-02
VIM	*LGG*	0.526	2.40e-08	7.01e-07
TUBB3	**KIRC**	0.344	4.00e-05	4.28e-04
FN1	*LGG*	0.309	1.20e-03	4.06e-03
	KIRP	0.484	1.30e-03	1.21e-02
	BLCA	0.294	5.70e-04	2.90e-02
RPL12	**KIRC**	0.304	1.70e-04	1.29e-03
	LGG	-0.320	1.50e-03	4.89e-03

KIRP: Kidney renal papillary cell carcinoma, LIHC: Liver Hepatocellular Carcinoma, LGG: Lower Grade Glioma, KIRC: Kidney Renal Clear Cell Carcinom, SARC: Sarcoma, CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma, PAAD: Pancreatic Adenocarcinoma, BLCA: Bladder Urothelial Carcinoma, LUAD: Lung Adenocarcinoma

p. E696 (2016).

[9] Taguchi, Y. H., Iwadate, M. and Umeyama, H.: SFRP1 is a possible candidate for epigenetic therapy in non-small cell lung cancer, *BMC Med Genomics*, Vol. 9 Suppl 1, p. 28 (2016).

[10] Taguchi, Y. H., Iwadate, M. and Umeyama, H.: Heuristic principal component analysis-based unsupervised feature extraction and its application to gene expression analysis of amyotrophic lateral sclerosis data sets, *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, pp. 1–10 (online), DOI: 10.1109/CIBCB.2015.7300274 (2015).

[11] Taguchi, Y. H., Iwadate, M. and Umeyama, H.: Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for post-traumatic stress disorder-mediated heart disease, *BMC Bioinformatics*, Vol. 16, p. 139 (2015).

[12] Taguchi, Y. H.: Principal component analysis based unsupervised feature extraction applied to budding yeast temporally periodic gene expression, *BioData Min*, Vol. 9, p. 22 (2016).

[13] Taguchi, Y. H.: Identification of aberrant gene expression associated with aberrant promoter methylation in primordial germ cells between E13 and E16 rat F3 generation vinclozolin lineage, *BMC Bioinformatics*, Vol. 16 Suppl 18, p. S16 (2015).

[14] 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* (Huang, D.-S., Han, K. and Gromiha, M., eds.), LNCS, Vol. 8590, Springer International Publishing, Heidelberg, pp. 445–455 (2014).

[15] Taguchi, Y. H., Iwadate, M., Umeyama, H., Murakami, Y. and Okamoto, A.: Heuristic principal component analysis-aased unsupervised feature extraction and its application to bioinformatics, *Big Data Analytics in Bioinformatics and Healthcare* (Wang, B., Li, R. and Perrizo, W., eds.), pp. 138–162 (2015).

[16] 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*, Vol. 15 Suppl 9, p. S2 (2014).

[17] Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., Ikeda, K., Kawada, N., Ochiya, T. and Taguchi, Y. H.: Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma, *Sci Rep*, Vol. 5, p. 16294 (2015).

[18] 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).

[19] 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).

[20] 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).

[21] 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).
[22] 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).
[23] 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).
[24] Taguchi, Y. H. and Okamoto, A.: Principal Component Analysis for Bacterial Proteomic Analysis, *Pattern Recognition in Bioinformatics* (Shibuya, T., Kashima, H., Sese, J. and Ahmad, S., eds.), LNCS, Vol. 7632, Springer International Publishing, Heidelberg, pp. 141–152 (2012).
[25] C.M. ビショップ：パターン認識と機械学習 上，丸善出版 (2012).
[26] Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 289–300 (online), available from ⟨http://www.jstor.org/stable/2346101⟩ (1995).
[27] Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., Tsai, W. T., Chen, G. Z., Lee, C. J., Chiu, C. M., Chien, C. H., Wu, M. C., Huang, C. Y., Tsou, A. P. and Huang, H. D.: miRTarBase: a database curates experimentally validated microRNA-target interactions, *Nucleic Acids Res.*, Vol. 39, No. Database issue, pp. D163–169 (2011).
[28] Anaya, J.: OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs, *PeerJ Computer Science*, Vol. 2, p. e67 (online), DOI: 10.7717/peerj-cs.67 (2016).