

主成分分析を用いた教師なし学習による変数選択法を用いたがんにおける mRNA-miRNA 相互作用のより信頼性のある同定

田口 善弘^{1,a)}

概要: microRNA (miRNA) は非常に多数の mRNA を標的とし、その標的の発現を抑止する機能を持つ転写後制御因子として注目を集めている。また、miRNA は発生、発達、各種疾患など広範な生物現象に関係していることが知られており、応用上も治療・診断標的として重要である。しかし、個々の miRNA がどのような状況でどの mRNA の発現制御に関わっているか(状況依存性)を網羅的に調べることは実験的に難しい。広く普及しているバイオインフォを使った予測は配列ベースであり、miRNA-mRNA 相互作用の状況依存性を反映できない。近年、このような限界を、miRNA/mRNA の発現プロファイルを用いることで乗り越え、状況依存性を考慮に入れた上で、miRNA-mRNA 相互作用を予測しようとする試みが盛んになっている。これらの研究では一般に miRNA-mRNA 相互作用の同定に先立って有意に発現が変化している miRNA/mRNA の同定を行うことが一般的であるが、有意に発現が変化している miRNA/mRNA の同定に大きな恣意性が見られることが多く、この結果、「適当な数の有意な miRNA-mRNA 相互作用を得られるように miRNA/mRNA の発現変化に関する有意性のしきい値をコントロールする」という本末転倒な事態になりがちである。本研究ではこの点を改善して、複数のがんの発現プロファイルに対して単一の基準で「有意な miRNA、mRNA 発現変化」を同定し、「有意性の基準」から「恣意性」を排除するために、著者が近年提案している教師なし学習を用いた主成分分析による変数選択法を用い、一定の成果を得た [1] のでここに報告する。

1. はじめに

miRNA の標的 mRNA は状況依存的であり、疾患・発生段階・時間など多岐にわたる要因に依存している。特に miRNA による遺伝子発現制御プロセスは転写後制御であり、転写プロセスが不変であっても最終的な翻訳生成物であるたんぱく質の生産量を変化させるという意味で、重層的なプロセスである。さらに、個々の miRNA が多数個(往々にして数百個以上)の mRNA を標的にする「可能性」を持っているという事実が状況を困難理解にする。一般にバイオインフォで予測できる miRNA-mRNA 相互作用は配列情報ベースでありこの「可能性」の予測に過ぎない(実際には、それさえ精度に限界があることが知られている)。一方、実験的にこの相互作用を検証するには膨大な資源を要するために現実的ではない。miRNA-mRNA 結合を固定して取り出した上に、その結合に含まれる miRNA 断片がどの miRNA 由来で、また、mRNA 断片がどの転写物

由来かを特定しなくてはならない。勢い、必要な配列断片の数も飛躍的に増大する。一度計測してしまえば、状況依存なく使用できる DNA 配列の解析や、mRNA や miRNA の断片の同定だけ行えばいい、RNA-seq や miRNA-seq に比べて手間がかかる割に、得られた情報は状況依存的であり、条件が変われば同じ miRNA-mRNA 結合を形成するという保証もない。限られた研究資源でなんらかの結果を出そうとすればこのような研究は避けられるのが理の当然である。勢い、十分な情報が得られない。

このような困難をわずかでも改善しようと、miRNA/mRNA の発現プロファイルの情報を用いて miRNA の標的 mRNA を特定しようという試みはかなり前から行われている。様々な試行錯誤を経て、現在のデファクトスタンダードは以下のようにになっている。まず、可能なすべての i, k ペアを考えるのではなく、mRNA, miRNA ごとに、操作群(疾患など)とコントロール群(健康者など)との間で有意な発現差があるものを選択することで考慮すべき i, k ペアの数を刈りこんでから、

$$x_{ij}^{\text{mRNA}} = a_{ik} x_{kj}^{\text{miRNA}} + b_i, j = 1, \dots, M, \quad (1)$$

¹ 中央大学理工学部物理学科 Tokyo 112-8551, Japan

^{a)} tag@granular.com

本研究内容はすでに原著論文として刊行済み [1] である

(1) 式の回帰分析を行う。ここで x_{ij}^{mRNA} は j 番目のサンプルにおける i 番目の mRNA の発現量、 x_{kj}^{miRNA} は j 番目のサンプルにおける k 番目の miRNA の発現量、 a_{ik}, b_i は回帰係数であり、 M はサンプルの総数である。(1) 式の回帰分析で有意な相関が見られた場合に、 i, k ペアが有意な miRNA-mRNA 相互作用であると見なす。この結果、考慮しなくてはならない i, k ペアの数減らすことができるようになり、miRNA-mRNA 相互作用の予測を miRNA/mRNA 発現プロファイルを用いて行うことが可能になり、多数の論文が出版されている。

一見、これで問題は解決したように見えたが、別の問題が起きているように思われる。操作群とコントロール群の間で有意に発現差がある miRNA, mRNA を選択する場合に任意性が残ってしまった。miRNA-mRNA 相互作用を発現プロファイルから同定する場合に、ちょうどいい miRNA, mRNA の数というのが存在する。あまり多くの miRNA, mRNA を残したのでは、 i, k ペアの数減らすという当初の目的が果たされない。一方、miRNA, mRNA の数を絞り込みすぎるとそもそもバイオインフォで予測された miRNA-mRNA のペアが少なくなりすぎてしまい、結果、同定できる miRNA-mRNA ペアの数ほとんどなくなってしまう。しかし、操作群とコントロール群の間で有意に発現差がある miRNA, mRNA の基準は M の数が増えれば、変化し、一般に大きな M ではごく僅かな差が有意と判定されてしまい、膨大な数の miRNA, mRNA が残ってしまう。このような場合にはいわゆるフォールド差でさらに絞り込みを行うことになるが、フォールド差による選別は、しきい値の設定に更に輪をかけて大きな自由度がある。この結果、操作群とコントロール群の間で有意に発現差がある miRNA, mRNA を選択する基準が、論文の執筆にちょうどよい個数の miRNA, mRNA を残すためのチューニングパラメータのようになってしまっている(表 1)。これでは生物学的に安定した結果を得るのは難しく、実験ごとに異なった結果になってしまう可能性を払拭できない。そこで今回我々は最近提唱している教師なし学習を用いた主成分分析による変数選択法 [13–26] を用いて複数のがんのプロファイルにおいて統一の基準で有意な発現変化がある miRNA/mRNA の選択を行うことができないかを試みた。その結果、それなりの成功を得たのでここに報告する [1]。

2. 結果

研究全体の流れを図 1 に描いた。本研究内容はすでに原著論文として刊行済みである [1] ので、本文中に不足している情報(例えば、同定された miRNA-mRNA ペアの具体的なリストなどもこの研究報告中には含めなかった)については原著論文を参照されたい。

表 2 に教師なし学習を用いた主成分分析による変数選択法(付録 A.1)で有意な発現差がある miRNA/mRNA の

表 1 有意な発現変化を持つ miRNA/mRNA の選定基準の一部、先行研究における例。

癌腫	有意差判別基準		文献
	miRNA	mRNA	
HCC	FDR ≤ 0.01 ; $\log_2 FC \geq 1$		[2]
NSCLC	FDR < 0.1 by SAM		[3]
ESCC	先行文献より	FC > 1.5	[4]
	FDR < 0.05	FC > 3 ; FDR < 0.001	[5]
	FDR < 0.05		[6]
PC	無し		[7]
CRC	FDR < 0.05		[8]
CC	FC > 1.2 ; FDR < 0.1		[9]
BC	miRtest [10]	記述なし	[11]
PDA	FDR* < 0.05 ; $ \log FC > 1$		[12]

HCC:肝細胞癌, NSCLC:非小細胞肺癌, ESCC:食道扁平上皮癌, PC:前立腺癌, CRC:大腸癌, CC:結腸癌, BC:乳癌, PDA:膵管癌, FC:fold change, *: Bonferroni's correction-adjusted p value

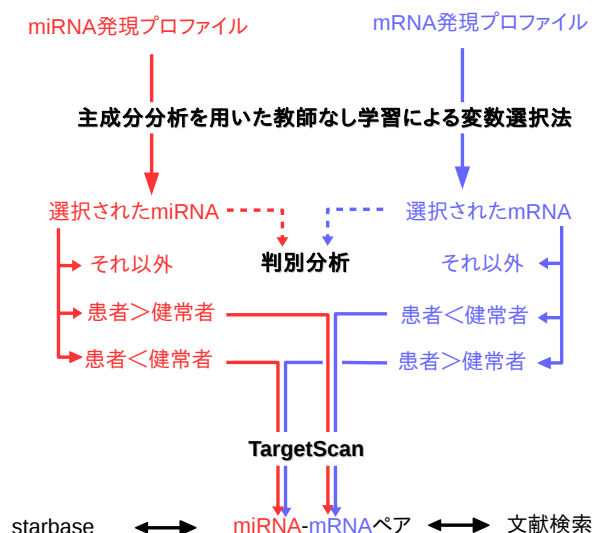


図 1 研究全体の流れ。

選択結果をまとめた (BH 基準 [27] で多重比較補正した P 値が 0.01 以下)。これらはそれぞれが異なったアレイ環境による計測(一部 NGS)であり、癌腫も多岐にわたっているし、サンプル数もまちまちだが、にも関わらず、有意な発現差があると同定された miRNA/mRNA の数は常識の範囲にとどまっている。生物学的な観点からして、がん化に過半数の miRNA/mRNA が関係しているとは思えないし、かといって、1 %を大きく割り込むというのも少なすぎる気がする。その意味では生物学的に妥当と想定される範囲の数が統一基準でうまく選定されたと言えるのではないだろうか。

次に、これらの選択された miRNA/mRNA がきちんとがん患者と健常者の差を表しているかを確認するために、選択された miRNA/mRNA のみを用いたがん患者/健常

表 2 有意に発現差がある mRNA/miRNA の選択結果まとめ。教師なし学習を用いた主成分分析による変数選択法で選択されかどうかを記述してある。

癌腫	GEO ID	サンプル数		プローブ数	
		腫患者	健常者	選択	非選択
HCC					
mRNA	GSE45114	24	25	269	22963
miRNA	GSE36915	68	21	58	1087
NSCLC					
mRNA	GSE18842	46	45	1098	53504
miRNA	GSE15008	187	174	268	3428
ESCC					
mRNA	GSE38129	30	30	189	22088
miRNA	GSE19337	76	76	37	1217
PC					
mRNA	GSE21032	150	29	399	43020
miRNA	GSE84318	27	27	23	700
CRC/CC					
mRNA	GSE41258	186	54	309	21974
miRNA	GSE48267	30	30	12	839
BC					
mRNA	GSE29174	110	11	980	33600
miRNA	GSE28884	173	16	18	2258

者の判別能を検証した(表 3)。方法は我々が教師なし学習を用いた主成分分析による変数選択法と組み合わせて使うことを提唱している「選択された miRNA/mRNA だけを用いて再度、miRNA/mRNA を主成分分析で低次元に埋め込み、得られた主成分負荷量(サンプル依存性を表現)を用いて線形判別を行う」という方法を採用した。判別率 100%である mRNA を用いた場合の非小細胞肺癌の判別をはじめとして、総じて判別能は良好である。このことから今回選択された miRNA/mRNA はきちんと患者と健常者の差を反映していると思われる。

最後に選択された miRNA/miRNA の間に TargetScan [28] で推定される標的関係がどれくらい含まれているかを確認した(conserved target のみを使用)。TargetScan は数多ある配列情報に基づく miRNA-mRNA 標的関係推定データベースの中でももっとも擬陽性が少ないデータベースとして知られており、特に conserved target に限ると擬陽性はかなり少なくなる。ただ、その場合、可能な数が限られてしまうので、先行研究で TargetScan の conserved set だけを用いたものは少ない。下手をするとヒット数が非常に減ってしまうかもしれないからである。しかし、今回の解析では、すべての癌腫において、mRNA,miRNA の発現変化が有意かつ逆方向(mRNA が有意に発現増大なら miRNA は有意に発現減少、あるいは、その逆の組み合わせ)であるものを 1 つ以上見つけることができた(表 4)。食道扁平上皮癌と大腸癌/結腸癌の場合、特定できた miRNA-mRNA 相互作用の数は数個になってしまったが、ゼロではない。更に、これらのペア

表 3 選択された miRNA/mRNA のみを用いた場合のがん患者/健常者の線形判別結果(混同行列)。主成分数: 判別に用いた主成分数。P 値とオッズ比はフィッシャーの正確確率検定で計算した。*は $P < 2.22E-16$ のもの。列が現実で行が予想。

	mRNA		miRNA	
	HCC	健常者	HCC	健常者
HCC	20	0	64	0
健常者	4	25	4	21
	(4, 3.75E-10, ∞)		(10, *, ∞)	
NSCLC	46	0	171	12
健常者	0	45	16	162
	(2, *, ∞)		(5, *, 1.39E+2)	
ESCC	28	2	63	11
健常者	2	28	13	65
	(2, 3.22E-12, 1.54E+2)		(6, *, 2.77E+1)	
PC	139	4	22	3
健常者	11	25	5	24
	(8, *, 7.45E+1)		(4, 2.88E-7, 3.17E+1)	
CRC/CC	178	5	27	3
健常者	8	49	3	27
	(8, *, 2.02E+2)		(4, 2.82E-10, 6.98E+1)	
BC	110	0	169	5
健常者	0	11	4	11
	(3, 7.83E-16, ∞)		(18, 2.62E-11, 8.49E+1)	

括弧内の数値は順に(主成分数, P 値, オッズ比)を表す。

に含まれる miRNA,mRNA の全てについて文献検索を行い、対象となるがんとの関連を報告する研究報告があるかを調べた。その結果、前立腺癌の場合を除き、同定された miRNA-mRNA ペアを構成する miRNA の全て、mRNA の大多数について対象となる癌腫との関係を示唆する既報が少なくとも一報はあった。前立腺癌のみなぜ既報のある mRNA が半数以下なのかの理由は不明だが、今回解析対象とした癌腫のうちでは前立腺癌だけが圧倒的に致死率が少ない(5 年生存率はほぼ 100%である)ため、そもそも、あまり研究対象となっていないのではないかと我々は考えている。

さらにペアについて starbase [29] において有意な miRNA-mRNA の負相関が報告されている miRNA-mRNA のペアの数を計数した。starbase では既存データベースから 14 種の癌の miRNA,mRNA のプロファイルをダウンロードして、負相関がある miRNA-mRNA をリスト化している。今回扱った 6 種類の癌が全て starbase に含まれているわけではないので、比較のため 14 種類の癌のうち 1 つでも負相関が確認できる miRNA-mRNA ペアをカウントした。表 4 にあるように大まかに言って半分弱のペアが

表 4 TargetScan の結果を用いた miRNA-mRNA 標的関係同定結果。ペアは miRNA-mRNA ペアのうち、TargetScan に含まれている数。miRNA,mRNA はそのペアを構成している miRNA,mRNA の数である。ペアの数より、miRNA,mRNA の数が少ないのは複数のペアに同一の miRNA,mRNA が含まれる場合があるからである。括弧内の数は miRNA,mRNA については、癌腫についての関連研究がある既報がある miRNA,mRNA の数、ペアについて starbase で有意な負相関が報告されているペアの数である。

癌腫	ペア		miRNA		mRNA	
HCC	20	(9)	13	(13)	18	(16)
NSCLC	311	(184)	27	(27)	113	(72)
ESCC	4	(2)	3	(3)	4	(4)
PC	32	(18)	8	(8)	19	(6)
CRC/CC	8	(3)	7	(7)	7	(6)
BC	37	(17)	11	(11)	30	(25)

starbase でなんらかの癌腫で有意な負相関を報告されていることから、今回の解析手法で得られた miRNA-mRNA ペアは実験とよく一致していると予想される。

3. 議論と結論

本研究では同定された miRNA-mRNA ペアの約半分が、独立な実験結果についても、負相関をもっているという結果を得た。これがどの程度の性能なのか定かではないが、TargetScan 単体の場合、負相関が観測されたのは 676265 ペアのうちのわずか 3210 ペア (0.5%以下) という報告もあるので [30]、予測性能が著しく向上したのは間違いないと思われる。

また、本研究では miRNA,mRNA で異ったコホートを使った上で、それぞれでがん患者-健康者間の発現変化が逆である miRNA-mRNA ペアを選択するという方法をとっている。異ったコホートを採用することで mRNA,miRNA 発現プロファイルの選択の自由度が広がることになる。異ったコホートを使っても精度の高い miRNA-mRNA 相互作用の同定ができるというのは本手法の合理的な点ではあるのであえてここで強調しておく。

一方で、他の先行研究との比較は未遂である。なぜなら、そもそも、配列情報に基づく miRNA-mRNA 予測と miRNA,mRNA の発現プロファイルの情報を統合的に解析した場合、それがどの程度の性能なのか、というチェックがなされている論文はなかなか無いからである。通常、この手の研究は目の前に miRNA,mRNA 発現プロファイルがあり、その中でどの miRNA がどの mRNA を標的としているかを知りたい、という動機でなされるので、出た結果を更に第三者のデータベースと突き合わせる、ということとはあまりされない。興味は目の前のデータにしかないので、結果が生物学的に妥当であれば、それ以上の探求はなされないのだ。一般にこのような動機でなされる研究が多い、ということが実際、表 1 に見られるように、miRNA,mRNA

の有意な発現差の同定が研究ごとにバラバラでも誰も気にしない、ということにつながっているのだろう。その意味では同定された miRNA-mRNA ペアがどの程度正しいのか、という問題意識自体、まだまだ少ないのでそもそも、本研究のようなことはなされないのかもしれない。この研究の評価も将来を待たなくてはならないだろう。

また、この手の研究(同定された miRNA-mRNA ペアの正確さの探求)が少ないのは、あまりにも多くのペアが同定されてしまうことが多くてチェックしきれない、ということも影響していると思われる。同定される miRNA-mRNA ペアの数数千に及ぶことが通常で、本研究にあるように、一個一個の miRNA や mRNA の文献検索や、個々のペアの starbase との比較などとてもやっていられない、というのが現実である。その意味では多重比較補正された P 値が 0.01 以下というごく当たり前の基準でありながら、同定されるペアの数を数十から数百程度に抑えることができる本手法は非常に有用な手法であると言えるだろう。

現在、miRNA-mRNA ペアを実験的に確認するにはシード領域と呼ばれる、miRNA が mRNA の標的とする領域に変異を入れて miRNA と mRNA の間の負相関が消失するかどうかを見る以外に手がない。非常に手間のかかる実験で、極端なことをいえば、一個のペアの実験的な検証ごとに論文が一本書かれているような状況である。したがって、本研究で得られたような miRNA-mRNA ペアを実験的に求めるのは相当先になる。それより、むしろ、今回これだけの数に絞れた miRNA-mRNA ペアの実験的な検証をぜひやりたいところだが残念ながら我々にはそれだけの設備がない。今後の研究を待ちたい。

最後に、最近 P 値を使った研究に対して、非常に風当たりが強いのだが、本研究のように P 値を計算する対象を工夫すればよいのであって、いたずらにベイズ統計や疎性モデリングの様な流行りの手法に走ってしまうのは、私見としてはどうかと思う。本研究では主成分分析という計算量的にはほとんどゼロに等しい計算しかしていないが、それでもこれだけの結果は出せる。計算機ではなく、頭を使うべきだ、と言ったら言いすぎだろうか？

参考文献

- [1] 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, p. 696 (2016).
- [2] Ding, M., Li, J., Yu, Y., Liu, H., Yan, Z., Wang, J. and Qian, Q.: Integrated analysis of miRNA, gene, and pathway regulatory networks in hepatic cancer stem cells, *J Transl Med*, Vol. 13, p. 259 (2015).
- [3] Ma, L., Huang, Y., Zhu, W., Zhou, S., Zhou, J., Zeng, F., Liu, X., Zhang, Y. and Yu, J.: An integrated analysis of miRNA and mRNA expressions in non-small cell lung cancers, *PLoS ONE*, Vol. 6, No. 10, p. e26502 (2011).

- [4] Wu, B., Li, C., Zhang, P., Yao, Q., Wu, J., Han, J., Liao, L., Xu, Y., Lin, R., Xiao, D., Xu, L., Li, E. and Li, X.: Dissection of miRNA-miRNA interaction in esophageal squamous cell carcinoma, *PLoS ONE*, Vol. 8, No. 9, p. e73191 (2013).
- [5] Yang, Y., Li, D., Yang, Y. and Jiang, G.: An integrated analysis of the effects of microRNA and mRNA on esophageal squamous cell carcinoma, *Mol Med Rep*, Vol. 12, No. 1, pp. 945–952 (2015).
- [6] Meng, X. R., Lu, P., Mei, J. Z., Liu, G. J. and Fan, Q. X.: Expression analysis of miRNA and target mRNAs in esophageal cancer, *Braz. J. Med. Biol. Res.*, Vol. 47, No. 9, pp. 811–817 (2014).
- [7] Zhang, W., Edwards, A., Fan, W., Flemington, E. K. and Zhang, K.: miRNA-mRNA correlation-network modules in human prostate cancer and the differences between primary and metastatic tumor subtypes, *PLoS ONE*, Vol. 7, No. 6, p. e40130 (2012).
- [8] Fu, J., Tang, W., Du, P., Wang, G., Chen, W., Li, J., Zhu, Y., Gao, J. and Cui, L.: Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis, *BMC Syst Biol*, Vol. 6, p. 68 (2012).
- [9] Li, X., Gill, R., Cooper, N. G., Yoo, J. K. and Datta, S.: Modeling microRNA-mRNA interactions using PLS regression in human colon cancer, *BMC Med Genomics*, Vol. 4, p. 44 (2011).
- [10] Artmann, S., Jung, K., Bleckmann, A. and Beissbarth, T.: Detection of simultaneous group effects in microRNA expression and related target gene sets, *PLoS ONE*, Vol. 7, No. 6, p. e38365 (2012).
- [11] Bleckmann, A., Leha, A., Artmann, S., Menck, K., Salinas-Riester, G., Binder, C., Pukrop, T., Beissbarth, T. and Klemm, F.: Integrated miRNA and mRNA profiling of tumor-educated macrophages identifies prognostic subgroups in estrogen receptor-positive breast cancer, *Mol Oncol*, Vol. 9, No. 1, pp. 155–166 (2015).
- [12] Liu, P. F., Jiang, W. H., Han, Y. T., He, L. F., Zhang, H. L. and Ren, H.: Integrated microRNA-mRNA analysis of pancreatic ductal adenocarcinoma, *Genet. Mol. Res.*, Vol. 14, No. 3, pp. 10288–10297 (2015).
- [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-based 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] 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).
- [17] 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).
- [18] 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).
- [19] 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).
- [20] 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).
- [21] 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).
- [22] 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).
- [23] 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).
- [24] 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).
- [25] 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).
- [26] 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).
- [27] 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).
- [28] Agarwal, V., Bell, G. W., Nam, J. W. and Bartel, D. P.: Predicting effective microRNA target sites in mammalian mRNAs, *Elife*, Vol. 4 (2015).
- [29] Li, J. H., Liu, S., Zhou, H., Qu, L. H. and Yang, J. H.: starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data, *Nucleic Acids Res.*, Vol. 42,

No. Database issue, pp. D92-97 (2014).

- [30] Wang, Y. P. and Li, K. B.: Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data, *BMC Genomics*, Vol. 10, p. 218 (2009).

付 録

A.1 主成分分析を用いた教師なし学習による変数選択法概要

詳しくはすでに刊行済みの原著論文 [13-26] を参照して頂きたいがここに概要 (図 A-1) を書いておく。

主成分分析を用いた教師なし学習による変数選択法においては通常の主成分分析と異なり、サンプルでなく遺伝子 (この場合は、miRNA, mRNA) を主成分分析で低次元に埋め込む。したがって、各遺伝子に主成分得点が付与される。この結果、主成分負荷量はサンプルに付与される。次に、変数選択に用いる主成分を選択する。いろいろな方法があるが、本研究の場合は単純に患者-健康者間で主成分負荷量に有意に差がある主成分を選んだ。次に、選択された主成分に含まれる主成分得点を用いて遺伝子の選択を行う。遺伝子の選択に置いては、主成分得点が多重ガウス分布をしているという帰無仮説 (遺伝子発現プロファイルが全くの乱数であれば主成分得点は中心極限定理からガウス分布に従うと期待される) に基づき、 χ^2 乗分布を仮定して、各遺伝子に P 値を付与した上、多重比較補正を施した

上で、外れ値 (多重ガウス分布に従うとみなせないもの、具体的には本研究では補正された P 値が 0.01 以下) の遺伝子を選択した。この操作は miRNA, mRNA 別々に行った。より詳しくは原著論文 [1] を参照して頂きたい。

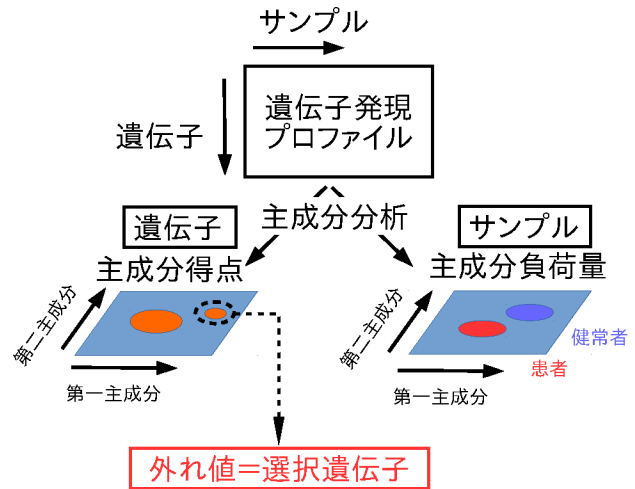


図 A-1 主成分分析を用いた教師なし学習による変数選択法概要。遺伝子発現プロファイル行列を主成分分析し、遺伝子に主成分得点、サンプルに主成分負荷量を割り当てる。主成分負荷量に患者と健康者とで差がある主成分 (図の例では第一、第二主成分) を特定し、対応する主成分得点で外れ値になっている遺伝子を特定し、これを選択遺伝子とする。