Vol.2013-BIO-33 No.8
2013/3/21

デジタルクラスタリングを用いた発がん過程の経時的トランスクリプトーム解析

青戸 良賢[1,a)]　八谷 剛史[3]　奥村 和弘[2]　長谷 純崇[1]　佐藤 健吾[1]　若林 雄一[2]　榊原 康文[1]

概要：発がんマウス実験系を用いて採取した正常上皮，良性腫瘍，悪性腫瘍，転移性腫瘍の 4 ステージに対して Illumina Genome Analyzer IIx による mRNA-Seq を行い，Student's t-test を用いて各ステージにおける発現差異遺伝子を探索した．本研究では，探索された発現差異遺伝子群の経時的トランスクリプトーム動態を解明するため，デジタルクラスタリングという手法を開発した．4 ステージ間計 6（$= {}_4C_2$）通りの検定結果をデジタル化した 6 次元ベクトルを構築し，これをマンハッタン距離を用いたウォード法による階層的クラスタリングを行うことで，同じ検定結果を有する遺伝子群を同一のクラスタに分類することができ，またクラスタ間で共通の検定結果を優先した階層的クラスタリングを行うことができる．本研究で得られた発現差異遺伝子候補群に対し既存のクラスタリング手法と比較を行った結果，既存手法では得ることができない，検定結果に基づいたクラスタが得られた．

キーワード：がん，経時的トランスクリプトーム解析，階層的クラスタリング

Temporal transcriptome analysis for carcinogenesis process by digital clustering.

Abstract: We developed the so called "digital clustering" method for analysis of temporal transcriptome dynamics of differential expression genes in carcinogenesis process. We obtained normal skin, papilloma, carcinoma, and metastasis by experimental carcinogenesis to mice, and sequenced mRNA by Illumina Genome Analyzer IIx. For searching differentially expressed genes, the statistical test, Student's t-test, was applied to the expression levels for each genes. We digitalized the results of total 6 tests, all pairwise combinations from 4 stages constitute six dimensional vectors. Digital clustering adopts hierarchical clustering with ward's method, using Manhattan distance, and by using 6 dimentional vectors, it can not only classify genes that have same test results into same cluster but also give priority to the common test results between clusters. The experiments on differential expression genes detected our method showed that digital clustering can provide significant clusters based on the statistical test results not provided by other existing methods.

Keywords: Cancer, Temporal transcriptome analysis, Hierarchical clustering

1. はじめに

がんは 1981 年に日本人の主要な死因となり，その発症数は年々増加している [1]．がんには良性腫瘍，悪性腫瘍，転移性腫瘍といった過程が存在し，異常増殖能，浸潤能，転移能などの機能を段階的に獲得することで悪性化が進行する．この多段階的な悪性化の過程で，腫瘍細胞内ではゲノムに変異が蓄積されていくことが知られており，このゲノム変異こそが悪性化の原因であると考えられている．がん悪性化に関わるゲノム変異には，1 塩基対が変異する点変異，染色体の重複，欠失，転座などがある．これらのゲノム変異が，がん遺伝子の優性な機能獲得，あるいはがん抑制遺伝子の機能損失と結びつくことで，がんの悪性化が進行する [2]．このことから，がんはゲノムの病気であり，がんの疾患メカニズムの解明にはゲノムを対象とした研究

[1] 慶應義塾大学理工学部生命情報学科
Department of Biosciences and Informatics, Keio University
[2] 千葉県がんセンター
Chiba Cancer Center Research Institute
[3] 岩手医科大学
Iwate Medical University
[a)] aoyocchi@dna.bio.keio.ac.jp

が必要不可欠であると言える．

がんのゲノム解析により多くの悪性化に関連するゲノム変異が同定された一方，悪性化に伴って変化する遺伝子発現のトランスクリプトーム解析が行われるようになった．近年，がんのトランスクリプトーム解析により，乳がんの亜型間に発現パターンの異なる遺伝子群が同定された [3] 他，食道扁平上皮がんにおいて新規がん抑制遺伝子が同定される [4] など，がん関連遺伝子探索に広く用いられている．これらのことから，がん研究においてトランスクリプトーム解析はゲノム解析と並んで重要なアプローチであると言える．

加えて現在，国際がんゲノムコンソーシアムによる大規模ながんゲノム解析が行われている．国際がんゲノムコンソーシアムでは，臨床的，社会的に重要とされる 50 種類の異なる腫瘍型および亜型について，ゲノム，トランスクリプトーム，エピゲノムなど包括的な解析が行われている [5]．これにより，主要ながんの臨床的知見が数多く得られることが期待される一方，解析の対象が主に悪性腫瘍および転移性腫瘍であることから，正常細胞から転移性腫瘍までの発がん過程全体を網羅的に解析することは難しく，発がん過程の全容を明らかにするためには正常細胞から良性腫瘍，悪性腫瘍，転移性腫瘍へと進行していく過程を経時的に追従し，網羅的解析を行う必要がある．

次世代シークエンサーを用いた経時的トランスクリプトーム解析の先行研究としては，プラナリアを用いた発現量解析が挙げられる [6]．この研究では，プラナリアの頭部を切除してから 0 min をコントロールとして，30 min, 1 h, 2 h, ... , 72 h までの計 16 タイムポイントについて，頭部の再生が行われている切除面を採取することで，経時的トランスクリプトーム解析を行っている．各タイムポイントで得られた発現量データを 0 min コントロールと比較，検定することにより発現差異遺伝子群を取得し，これらの発現パターンを z スコアを用いたユークリッド距離による k-means クラスタリングにより分類することで共発現遺伝子を探索している．しかし，この手法による時系列クラスタリングでは，各クラスタに異なるステージ間で有意差が認められた遺伝子が混在してしまい，検定結果の反映されていないクラスタが生成されてしまう．

そこで本研究では，DMBA/TPA 処理による 2 段階発がんマウス実験を用いることにより，同一個体マウスから正常上皮，良性腫瘍，悪性腫瘍，転移性腫瘍を採取することで時系列サンプルを取得した．これら 4 つのステージから次世代シークエンサー（Illumina Genome Analyzer IIx）による mRNA-Seq により経時的トランスクリプトームを取得し，時系列データのクラスタリングを高精度に行うデジタルクラスタリングの手法を開発し，発がん過程に特徴的な遺伝子群の探索を行った．

2. 方法

2.1 発がん実験および腫瘍の採取

7,12-Dimethylbenz（a）anthracene（DMBA）/12-*O*-tetradecanoylphorbol-13-acetate（TPA）薬剤処理による発がん実験を 17 個体の FVB 系統マウスに対し行うことで，同一個体由来の正常上皮，良性腫瘍，悪性腫瘍，転移性腫瘍を採取した．DMBA を発がん物質，TPA をプロモーターとして用いることで，2 段階発がん実験を行った．背中の毛を電動剃刀で剃った 2 日後に，各マウスの背中に対し DMBA のアセトン希釈液（アセトン 200 μL に対し，DMBA 25 μL）を単回投与した．その 1 週間後より TPA のアセトン希釈液（アセトン 200 μL に対し，TPA 10 μL）を週 2 回，20 週間投与した．良性腫瘍と悪性腫瘍の区別は外観により決定し，背中皮膚より伸び出た腫瘍を良性腫瘍，真皮に浸潤し平らになったものを悪性腫瘍とした．この実験系を用いて，同一個体マウスから正常上皮，良性腫瘍，悪性腫瘍，転移性腫瘍の経時的サンプルを採取した．

2.2 RNA 抽出および mRNA-Seq

採取した 4 つの組織から AGPC 法 [7] により Total RNA を抽出した．ただしエタノール沈殿して乾燥させた後，Total RNA を DEPC 水に溶かすところを 10 μL DNase Buffer（QIAGEN 79254）に溶かし，1.4 μL DNase（QIAGEN 79254）を加えて室温放置することにより DNA の除去を行った．D-solution を 500 μL 加えた後，再び AGPC 法により抽出を行った．この 2 回目の AGPC 法において，D-solution で処理する過程は省略した．

得られた Total RNA の cDNA 合成を SMARTer^{TM} Ultra Low RNA Kit（Clontech 634935）を用いて行い，アコースティックゾルビライザー（Covaris inc.）により 500 bp への断片化を行った．PCR Purification kit（QIAGEN 28104）による精製後，Illumina 社の Paired-End Sample Preparation Guide に従ってサンプル調整を行った．正常上皮，良性腫瘍，悪性腫瘍，転移性腫瘍の各サンプルを 6 pM に希釈し，各サンプルにつき 1 レーンずつ Illumina Genome Analyzer IIx による Paired-end mRNA-Seq（60 nt）を行った．

2.3 デジタルクラスタリング

UCSC よりマウスゲノム（mm9）およびアノテーション情報を取得し，mRNA-Seq により得られたペアエンドリードをマウスゲノムに対し Bowtie2 [8] を用いてマッピングした．Bowtie2 によりマッピングすることができなかったリードのうち，N がなく，40 塩基以上で，Quality value が 20 以下の塩基を含まないリードを Tophat [9] によりシングルエンドリードとしてマッピングした．マッピング結

果を用いて Cuffdiff2 [10] による正規化発現 FPKM [11] の算出，および発現差異遺伝子探索（FDR:False Discovery Rate $< 5\%$）を行った．4 ステージ間計 6（$= {}_4C_2$）通りの検定を行い，2 つのステージ間で有意に発現量の異なる遺伝子を発現差異遺伝子候補とした．

得られた発現差異遺伝子候補群に対し，本研究で開発したデジタルクラスタリングを適用した．2 つのステージ間で有意に発現量が増加している場合には 1，減少している場合には -1，有意差がない場合には 0 として，すべての遺伝子ペア間の 6 通りある検定結果を，6 次元のベクトルにデジタル化した．この 6 次元デジタルベクトルに対し，マンハッタン距離を用いてウォード法 [12] による階層的クラスタリングを行った．

まず，遺伝子 A の 6 次元ベクトルを $x_A = (x_{A1}, x_{A2}, ..., x_{A6})$ とすると，遺伝子 A，B のマンハッタン距離 d_M は以下（1）式で定義される．

$$d_M(x_A, x_B) = \sum_{i=1}^{6} |x_{Ai} - x_{Bi}| \quad (1)$$

次に，ウォード法で階層的クラスタリングを行う．ウォード法とはクラスタ内の各構成要素から重心までの距離の総和 S が最小となるようにクラスタを形成する手法である．2 つの n 次元ベクトル x_A，x_B の距離を $d(x_A, x_B)$ とし，m 個の構成要素から成るクラスタ C_k の重心を $\bar{x}$ とすると，重心までの距離の総和 $S(C_k)$ は（2）式で定義される．

$$S(C_k) = \sum_{i=1}^{m} d(\bar{x}, x_i) \quad (2)$$

これを用いることで，2 つのクラスタ C_1，C_2 間の距離 $D(C_1, C_2)$ は（3）式で定義することができる．

$$D(C_1, C_2) = S(C_1 \cup C_2) - S(C_1) - S(C_2) \quad (3)$$

（3）式より，距離 D は 2 つのクラスタをマージする前後における，重心までの距離の総和 S の変化量 ΔS に一致する．このことから，距離 D が最小となるペアを逐次的にマージしていくウォード法では，クラスタ内の分散が最小となるように階層的クラスタリングを行うことができる．

そこで本研究では，ベクトル間の距離をマンハッタン距離で定義し，6 次元ベクトルのウォード法による階層的クラスタリングを行うことで，発がん過程における遺伝子発現パターンを高精度に分類した．まず，同一の検定結果を持つ遺伝子間のマンハッタン距離は 0 であることから，6 次元の各ベクトルに分類される遺伝子は，同一の検定結果を持つ遺伝子群である．以降，1 つの 6 次元ベクトルが表す検定結果を持つ遺伝子のクラスタを最小クラスタと呼ぶこととする．これにより，同一の検定結果を持つ遺伝子群を必ず 1 つのクラスタに分類することができる．

また，クラスタ内の重心を構成要素の平均として定義し，（2）式，（3）式に（1）式を代入することによりクラスタ間の距離を定義することで，ウォード法による階層的クラスタリングを行った．これにより，多くの共通の検定結果を持つクラスタ間の距離を近く定義することができ，それらを逐次的にマージすることにより，発がん過程において共通の発現パターンを有する遺伝子群を分類することができる．すなわち，本手法の適用により，発がん過程に特徴的な遺伝子群の探索が可能となる．

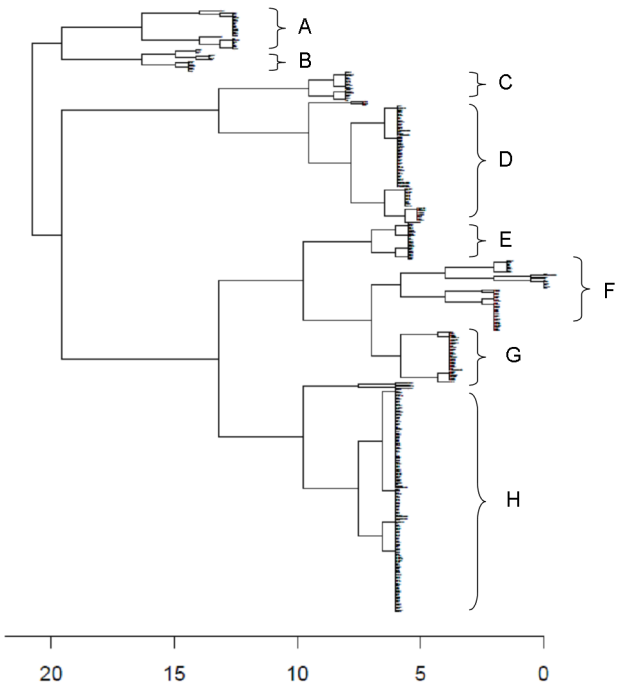

図 1　デジタルクラスタリングにより得られた系統樹．
Fig. 1　Dendrogram drawn by digital clustering.

3. 結果

3.1 デジタルクラスタリングの結果

2.3 節の発現差異遺伝子探索により 374 個の遺伝子を発現差異遺伝子候補として同定した．これらの遺伝子群に対し本研究で開発したデジタルクラスタリングを適用することで，図 1 の発現パターンによる系統樹が得られた．

まず，374 個の発現差異遺伝子候補は 48 種類の検定結果に分類することができた．また図 1 の右側に示したように，A から H まで計 8 つの特徴的なクラスタに分類することができた．図 2~5 は，クラスタ A，B，D，H の正規化発現量 FPKM を描いたグラフである．クラスタ A には正常上皮で発現量が低く，転移性腫瘍で発現量が高い 24 個の遺伝子が所属している．またクラスタ B には，良性腫瘍で発現量が低く，悪性腫瘍で発現量の高い 14 個の遺伝子が所属している．図 2，3 より，これら 2 つのクラスタに属する遺伝子は発現パターンの形状だけを見ると類似しているが，本手法では検定結果の差異により区別することができる．図 4 から分かるように，クラスタ D には良性腫瘍

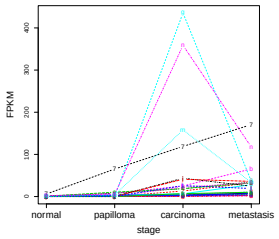

図 2　クラスタ A の FPKM.
Fig. 2　FPKM of cluster A.

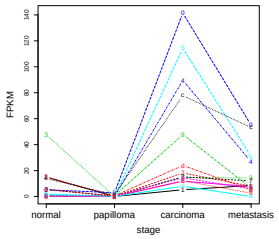

図 3　クラスタ B の FPKM.
Fig. 3　FPKM of cluster B.

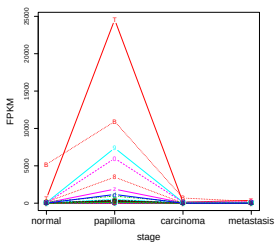

図 4　クラスタ D の FPKM.
Fig. 4　FPKM of cluster D.

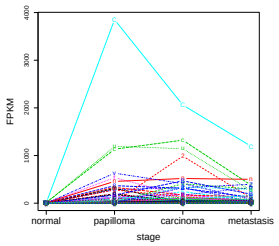

図 5　クラスタ H の FPKM.
Fig. 5　FPKM of cluster H.

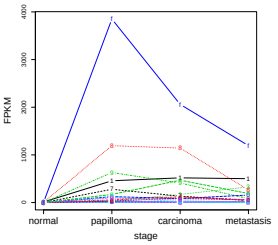

図 6　クラスタ H に含まれるクラスタ 2 の FPKM.
Fig. 6　FPKM values of cluster 2 derived cluster H.

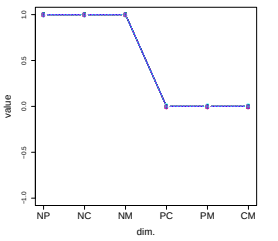

図 7　クラスタ H に含まれるクラスタ 2 の 6 次元ベクトル
Fig. 7　6 dim. vectors of cluster 2 derived cluster H.

において発現量が増加している遺伝子群が属している．またクラスタ H には正常上皮で発現量が低く，悪性腫瘍で発煙量の高い遺伝子群が属している．このように，図 5 からでは解釈することが難しいクラスタも検定結果を得ることによって解釈が可能となる．よって階層的クラスタリングを行うことにより，検定結果の異なる遺伝子群を区別するだけでなく，各発がん過程に特徴的な遺伝子群を得ることができる．

加えて，図 1 の系統樹におけるクラスタ H に含まれる，2 番クラスタの正規化発現量 FPKM および 6 次元ベクトルを描いたものが図 6，7 である．図 7 の横軸は，N を正常上皮，P を良性腫瘍，C を悪性腫瘍，M を転移性腫瘍として，2 つのステージ間における検定結果を表してる．このクラスタ 2 には 18 個の遺伝子が属しており，検定結果を表す 6 次元ベクトルから，正常上皮に対して良性腫瘍，悪性腫瘍および転移性腫瘍で発現量が高い遺伝子群である．この 18 個の遺伝子には，細胞周期の進行に関わる *Ccnd*1 [13] や，卵巣がんの診断マーカーとして報告されている *Klk*13 [14]，子宮内膜がんで細胞増殖に関与が疑われる *P*2*y*2 [15] などが含まれており，既知のがん関連遺伝子が腫瘍内で高発現となっていることが示唆された．

3.2　既存手法との比較

1 節で述べた通り，既存手法として，正規化発現量に対してユークリッド距離を用いた k-means クラスタリングを行うことで，遺伝子の発現パターンの分類を行っているものがある．そこで，本研究で同定した 374 個の発現差異遺伝子候補群に対し，正規化発現量をもとに，ユークリッド距離を用いた k-means クラスタリングを行った．3.1 節で述べた通り，374 個の遺伝子は 48 個の要素に分類できることから，$k = 48$ として行った．図 8~10 は k-means クラスタリングにより得られた 48 クラスタのうち，クラスタ 6，13，34 に割り当てられた遺伝子群の正規化発現量 FPKM を描いた図である．また表 1 はこれらのクラスタに属する遺伝子数と，それらの検定結果の種類数をまとめたものである．検定結果の種類とは 6 次元ベクトルの種類のことを指し，種類数とはクラスタ内の各遺伝子が持つ 6 次元ベクトルが，クラスタ内で合計何種類あるかを表す．つまり，遺伝子数に対して検定結果の種類が少ないほど，1 つのクラスタ内に同一の検定結果を持つ遺伝子が分類されていることを表す．

表 1　各クラスタの遺伝子数と，検定結果の種類.
Table 1　The number of genes and types of test results for each clusters.

クラスタ番号	遺伝子数	検定結果の種類
クラスタ 6	8	7
クラスタ 13	6	6
クラスタ 34	20	19

図 8~10 より，概ね発現パターンにより分類されていることが分かる一方，表 1 より，各クラスタに属する遺伝子はほとんどが異なる検定結果を持っており，同じ検定結果を持つ遺伝子群が同一のクラスタに分類されていないことが分かる．さらに，3.1 節で示した図 6，7 の 2 番クラスタに属する 18 個の遺伝子は，既存手法では，上記のクラスタ 6，13，34 を含めた計 17 個のクラスタに分類された．これ

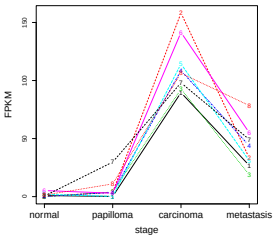

図 8 *k*-means クラスタリングの結果 (クラスタ 6).

Fig. 8 Result of *k*-means clustering (cluster 6).

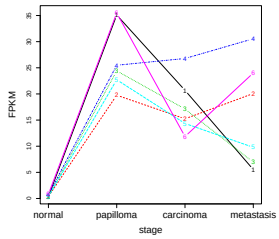

図 9 *k*-means クラスタリングの結果 (クラスタ 13).

Fig. 9 Result of *k*-means clustering (cluster 13).

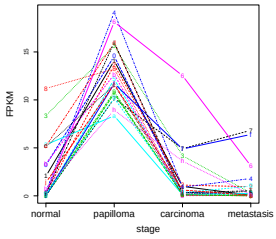

図 10 *k*-means クラスタリングの結果 (クラスタ 34).

Fig. 10 Result of *k*-means clustering (cluster 34).

らのことから，既存手法では検定結果を反映したクラスタに分類することができないことが示唆された一方，本手法では検定結果に基づいて高精度に分類することができる．

検定結果に基づいてクラスタリングを行う利点としては，各遺伝子の各ステージにおける発現量の分散を考慮できる点である．各ステージに複数のサンプルが用意された場合に，多くの既存手法では主に正規化発現量の平均値のみを考慮したクラスタリングが行われている．しかし，ある 2 つの遺伝子が同一のクラスタに分類され，発現パターンの形状が類似していたとしても，分散の大小によって有意差の有無に違いが生まれる．この有意差の有無に基づいて発現パターンを分類することにより，解釈可能な，意味のあるクラスタに分類することができる．本研究では各ステージにつきサンプル数が 1 つであるため，Cuffdiff2 による，推定平均および推定分散を用いた検定を行っている．

4. おわりに

本論文では，時系列トランスクリプトームデータの検定結果をデジタル化して階層的クラスタリングを行う，デジタルクラスタリングを提案した．正規化発現量をクラスタリングする既存手法では分類することができない，検定結果に基づくクラスタが得られ，これを階層的にマージすることで各発がん過程に特徴的な遺伝子群を探索した．

今後は良性腫瘍，悪性腫瘍を同一の腫瘍から採取した時系列サンプルの mRNA-Seq を行い，本手法を適用することで，がん悪性化に寄与する遺伝子の同定を行う．

参考文献

[1] Utada, M., Ohno, Y., Soda, M. and ichi Kamo, K.: Estimation of cancer incidence in Japan with an age-period-cohort model., *Asian Pac J Cancer Prev*, Vol. 11, No. 5, pp. 1235–1240 (2010).

[2] Hanahan, D. and Weinberg, R. A.: The hallmarks of cancer., *Cell*, Vol. 100, No. 1, pp. 57–70 (2000).

[3] Kao, K.-J., Chang, K.-M., Hsu, H.-C. and Huang, A. T.: Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization., *BMC Cancer*, Vol. 11, p. 143 (online), DOI: 10.1186/1471-2407-11-143 (2011).

[4] Tong, M., Chan, K. W., Bao, J. Y. J., Wong, K. Y., Chen, J.-N., Kwan, P. S., Tang, K. H., Fu, L., Qin, Y.-R., Lok, S., Guan, X.-Y. and Ma, S.: Rab25 is a tumor suppressor gene with antiangiogenic and anti-invasive activities in esophageal squamous cell carcinoma., *Cancer Res*, Vol. 72, No. 22, pp. 6024–6035 (online), DOI: 10.1158/0008-5472.CAN-12-1269 (2012).

[5] et al., I. C. G. C.: International network of cancer genome projects., *Nature*, Vol. 464, No. 7291, pp. 993–998 (online), DOI: 10.1038/nature08987 (2010).

[6] Sandmann, T., Vogg, M. C., Owlarn, S., Boutros, M. and Bartscherer, K.: The head-regeneration transcriptome of the planarian Schmidtea mediterranea., *Genome Biol*, Vol. 12, No. 8, p. R76 (online), DOI: 10.1186/gb-2011-12-8-r76 (2011).

[7] Chomczynski, P.: A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples., *Biotechniques*, Vol. 15, No. 3, pp. 532–4, 536–7 (1993).

[8] Langmead, B. and Salzberg, S. L.: Fast gapped-read alignment with Bowtie 2., *Nat Methods*, Vol. 9, No. 4, pp. 357–359 (online), DOI: 10.1038/nmeth.1923 (2012).

[9] Trapnell, C., Pachter, L. and Salzberg, S. L.: TopHat: discovering splice junctions with RNA-Seq., *Bioinformatics*, Vol. 25, No. 9, pp. 1105–1111 (online), DOI: 10.1093/bioinformatics/btp120 (2009).

[10] Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L. and Pachter, L.: Differential analysis of gene regulation at transcript resolution with RNA-seq., *Nat Biotechnol*, Vol. 31, No. 1, pp. 46–53 (online), DOI: 10.1038/nbt.2450 (2013).

[11] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation., *Nat Biotechnol*, Vol. 28, No. 5, pp. 511–515 (online), DOI: 10.1038/nbt.1621 (2010).

[12] Ward Jr, J. H.: Hierarchical grouping to optimize an objective function, *Journal of the American statistical association*, Vol. 58, No. 301, pp. 236–244 (1963).

[13] Bell, J. L., Malyukova, A., Kavallaris, M., Marshall, G. M. and Cheung, B. B.: TRIM16 inhibits neuroblastoma cell proliferation through cell cycle regulation and dynamic nuclear localization., *Cell Cycle*, Vol. 12, No. 6 (2013).

[14] Scorilas, A., Carla ABorgo 単 o, Harbeck, N., Dorn, J., Schmalfeldt, B., Schmitt, M., Diamandis, E. P. : Human kallikrein 13 protein in ovarian cancer cytosols: a new fa-

vorable prognostic marker., *J Clin Oncol*, Vol. 22, No. 4, pp. 678–685 (online), DOI: 10.1200/JCO.2004.05.144 (2004).

[15] Katzur, A. C., Koshimizu, T., Tomic, M., Schultze-Mosgau, A., Ortmann, O. and Stojilkovic, S. S.: Expression and responsiveness of P2Y2 receptors in human endometrial cancer cell lines., *J Clin Endocrinol Metab*, Vol. 84, No. 11, pp. 4085–4091 (1999).