Vol.2019-BIO-57 No.6
2019/3/9

主成分分析を用いた教師無し学習による変数選択の一細胞RNA-seqへの応用

田口善弘[1,a)]

概要：一細胞 RNA-seq は従来の臓器レベルの平均化された遺伝子発現プロファイルの観測を超えて、細胞ごとの発現プロファイルを観測できるという意味で画期的である。一方、個々の細胞にはラベルがついていないため、従来の臓器レベルの観測の様に、正常臓器と腫瘍で差が大きい遺伝子を選ぶ、などの基準で遺伝子を選択することができない。遺伝子を選択することは tSNE などのクラスタリングによる可視化を行う場合にも非常に重要なプロセスである。このため、ラベルを用いない教師なし学習による変数選択の方法がいくつか提案されてきた。ここでは、著者が従来から提唱している「主成分分析を用いた教師なし学習による変数選択法」を一細胞 RNA-seq における遺伝子選択に用いた場合を考察し、他の手法（highly variable genes, bimodal genes, dpFeature）による変数選択との比較を行う。

1. はじめに

一細胞 RNA-seq (single cell RNA-seq, 以下、scRNA-seq) は遺伝子発現プロファイルの観測手段として、従来の臓器ベースの観測に比べて、大きな利点をもっている。それは臓器内の細胞の多様性を観測できる、ということである。ガンの例に見るまでもなく、単一の臓器であっても、その内部は一様な構造ではなく、勢い、臓器ごとの遺伝子発現プロファイルの計測はその様な多様性を無視した、臓器全体で平均化された遺伝子発現プロファイルに過ぎない。極端な場合には、臓器単位で平均された遺伝子発現プロファイルを実際に呈している細胞は1 つもない、ということもあり得る。もし、そうなってしまっているとすると、得られた遺伝子発現プロファイルから、臓器特異的な発現を呈している遺伝子を選択し、エンリッチメント解析や、パスウェイ解析で生物学的な機序を得ようとすること自体が、実際はまったくの見当違いである、という可能性さえ考えられる。

scRNA-seq はこの様な問題を超えて、臓器内の細胞のごとの遺伝子発現プロファイルの多様性を把握できるポテンシャルを秘めている。一方で、腫瘍対正常臓器のようなわかりやすい対照群をもっていないため、scRNA-seq で特徴的に発現している遺伝子はどれか、という問に答えるのは簡単ではない。単に発現量が多くても、どの細胞でも同じように発現しているのであれば、そもそもそれはハウスキーピング遺伝子かもしれず、なんら研究対象と関係ないものかもしれない。かと言って、単純に発現差が大きい遺伝子を選んでも、それは大きなノイズが乗っているだけで生物学的には無意味かもしれない。この問題は典型的な教師なし学習による変数選択問題であり、従来から著者が提唱してきた「主成分分析を用いた教師なし学習による変数選択」の適用例として適当である。そこで本稿では同手法を scRNA-seq の遺伝子選択に用いる試みを紹介すると共に既存手法との比較を行う。

[1] 中央大学理工学部物理学科
[a)] tag@granular.com
本研究は国際会議 ICIC2018 で発表済み、プロシーディングスとして出版済みである [1, 2]。

2. データと方法

2.1 scRNA-seq 発現プロファイルデータ

解析対象の scRNA-seq 発現プロファイルデータ [3] は GEO の GSE76381 からダウンロードした。具体的には Supplementary file セクションで提供されている、GSE76381_EmbryoMolecule Counts.cef.txt.gz という名前のファイル（ヒト）と GSE76381_MouseEmbryoMolecule Counts.cef.txt.gz という名前のファイル（マウス）をダウンロードして用いた。これらは共に、腹側中脳の発生過程での遺伝子発現プロファイルであり、複数の時系列での計測から構成されている。以下の解析では遺伝子選択に際して時系列の情報は用いなかった。

2.2 主成分分析を用いた教師なし学習による変数選択を用いた遺伝子選択

まず i 番目の遺伝子の j 番目の細胞における発現量

$x_{ij} \in \mathbb{R}^{N\times M}$ を $\sum_i x_{ij} = 0, \sum_i x_{ij}^2 = N$ になるように標準化する。次に $S_{ii'} = \sum_j x_{ij}x_{i'j}$ で定義される行列を対角化し、固有ベクトル $\boldsymbol{u}_\ell \in \mathbb{R}^N$ を計算することで主成分得点を計算し、主成分分析を実行する。この $\boldsymbol{u}_\ell$ を用いて、$u_{\ell i}$ がガウス分布であるという帰無仮説のもとに、χ 二乗分布を使って i 番目の遺伝子に P 値 P_i を

$$P_i = P_{\chi^2}\left[> \sum_{\ell=1}^{L} \left(\frac{u_{\ell i}}{\sigma_\ell}\right)^2 \right]$$

という式で付与する。ここで σ_ℓ は標準偏差、$P_{\chi^2}[> x]$ は引数が x より大きい時の χ 二乗分布の累積確率である。BH 基準 [4] で多重比較補正した P 値が 0.01 以下の miRNA を選択する。

2.3 Highly variable genes を用いた遺伝子選択

Highly variable genes [5] のアルゴリズムは以下のとおりである。まずは遺伝子の発現量の平均 μ_i

$$\mu_i = \frac{\sum_{j=1}^{M} x_{ij}}{M}$$

と標準偏差 σ_i

$$\sigma_i^2 = \frac{\sum_{j=1}^{M}(x_{ij} - \mu_i)^2}{M}$$

を計算する。次に回帰式

$$\log_{10}\left(\frac{\sigma_i}{\mu_i}\right) = \frac{1}{2}\log_{10}\left(\frac{\beta}{\mu_i} + \alpha\right) + \epsilon_i$$

を用いて、回帰係数 α, β と残差 ϵ_i を計算する。そして、

$$P_i = P_{\chi^2}\left[> \left(\frac{\epsilon_i}{\sigma'}\right)^2 \right]$$

で計算した P 値を BH 基準で多重比較補正した P 値が 0.01 以下の遺伝子を選択する (σ' は標準偏差)。

2.4 Bimodal genes を用いた遺伝子選択

Unimodal test を R に実装されている dip.test 関数で各遺伝子に実行し、P 値を付与する。BH 基準で多重比較補正した P 値が 0.01 以下の遺伝子を選択する。

2.5 dpFeature を用いた遺伝子選択

dpFeature [6] を用いる。

3. 結果

まずは主成分分析を用いた教師なし学習による変数選択の結果を述べる。ヒトに対しては $L = 2$ で 116 遺伝子が、マウスに対しては $L = 3$ で 118 遺伝子が選択された。共通に選ばれた遺伝子は５３遺伝子もあった。偶然で、まったく別の遺伝子発現プロファイルからこれだけの数が共通に選ばれるとは考えにくい。しかし、一方で生物学とはまったく無関係な理由で選ばれている可能性もある。そこでこれらの遺伝子の生物学的な意義の評価を行った。表 **1** は Enrichr の “MGI Mammalian Phenotype 2017” のカテゴリによるマウスに対して同定された 118 遺伝子の評価である。上位5 位のうち、4 位までが脳の形態形成の異常を選択している。発生過程であるから、通常の組織の発現からすると以上と診断されたのであろうがちゃんと脳の形態形成の関連遺伝子がもっとも高い頻度で選ばれている。とりあえずは、選択はうまく言っているといっていいだろう。表 **2** はヒトに対して同定された 116 遺伝子の Enrichr の “Allen Brain Atlas down” の上位5 位の結果である。いずれも脳に関連している。やはり、発生過程なので、正常な脳に比べると発現していない脳に関連した遺伝子が発現しているということで妥当である。表 **3** はマウスに対して同定された 118 遺伝子の Enrichr の “Allen Brain Atlas down” の上位5 位の結果である。いずれも脳に関連している。やはり、発生過程なので、正常な脳に比べると発現していない脳に関連した遺伝子が発現しているということで妥当である。表 **4** はヒトに対して同定された 116 遺伝子の Enrichr の “GTEx Tissue Sample Gene Expression Profiles down” の上位5 位の結果である。いずれも脳に関連している。やはり、発生過程なので、正常な脳に比べると発現していない脳に関連した遺伝子が発現しているということで妥当である。表 **5** はマウスに対して同定された 118 遺伝子の Enrichr の “GTEx Tissue Sample Gene Expression Profiles down” の上位5 位の結果である。上位5 位中、2,4,5 位は脳に関連している。やはり、発生過程なので、正常な脳に比べると発現していない脳に関連した遺伝子が発現しているということで妥当である。ここまではみな、発現が下がっている遺伝子ばかりだった。上がっているものの妥当性を確認するため、同じ Enrichr の “Jensen TISSUES” の Embryonic_brain 結果をみてみた (表 **6**)。非常に高い有意度で胎児の脳で上昇している遺伝子が上昇している。その数はヒトとマウスがそれぞれ、７１遺伝子と７５遺伝子であり、ヒトとマウスに対して、主成分分析を用いた教師なし学習による変数選択でヒトとマウスについて選ばれた１１６遺伝子と１１８遺伝子のゆうに過半数を超えた遺伝子を占めている。これらのことから同手法の遺伝子選択は生物学的にみて極めて妥当であると考えられる。

さらに、選択された遺伝子から、発現制御機構に対するヒントが得られるのではないかと考えた。表 **7** は主成分分析を用いた教師なし学習による変数選択でヒトに対して選択された 116 遺伝子とマウスに対して選択された 118 遺伝子を Enrichr にアップロードしたときの “ENCODE and ChEA Consensus TFs from ChIP-X” の結果(adjusetd P-values が 0.01 以下の遺伝子) である。太字がマウスとヒトで共通に選ばれたものである。ヒトでは４２個の、マウス

表 1　主成分分析を用いた教師なし学習による変数選択でマウスに対して選択された 118 遺伝子を Enrichr にアップロードしたときの “MGI Mammalian Phenotype 2017” のカテゴリの上位5 位までの結果

Term	Overlap	P-value	Adjusted P-value
MP:0000788_abnormal_cerebral_cortex_morphology	7/145	2.45×10^{-5}	4.55×10^{-5}
MP:0003651_abnormal_axon_extension	5/48	9.18×10^{-6}	4.55×10^{-3}
MP:0000812_abnormal_dentate_gyrus_morphology	5/58	2.34×10^{-5}	4.55×10^{-3}
MP:0000807_abnormal_hippocampus_morphology	5/86	1.56×10^{-4}	2.04×10^{-2}
MP:0000819_abnormal_olfactory_bulb_morphology	4/48	1.83×10^{-4}	2.04×10^{-2}

表 2　主成分分析を用いた教師なし学習による変数選択でヒトに対して選択された 116 遺伝子を Enrichr にアップロードしたときの “Allen Brain Atlas down” のカテゴリの上位5 位までの結果

Term	Overlap	P-value	Adjusted P-value
periventricular stratum of cerebellar vermis	18/300	1.33×10^{-13}	3.72×10^{-11}
Simple lobule	18/300	1.33×10^{-13}	3.72×10^{-11}
Simple lobule, molecular layer	18/300	1.33×10^{-13}	3.72×10^{-11}
Simple lobule, granular layer	18/300	1.33×10^{-13}	3.72×10^{-11}
white matter of cerebellar vermis	18/300	1.33×10^{-13}	3.72×10^{-11}

表 3　主成分分析を用いた教師なし学習による変数選択でマウスに対して選択された 118 遺伝子を Enrichr にアップロードしたときの “Allen Brain Atlas down” のカテゴリの上位5 位までの結果

Term	Overlap	P-value	Adjusted P-value
Pyramus (VIII), granular layer	18/300	1.81×10^{-13}	4.66×10^{-11}
Pyramus (VIII)	18/300	1.81×10^{-13}	4.66×10^{-11}
Pyramus (VIII), molecular layer	18/300	1.81×10^{-13}	4.66×10^{-11}
Paraflocculus, molecular layer	18/300	1.81×10^{-13}	4.66×10^{-11}
Cerebellar cortex	18/300	1.81×10^{-13}	4.66×10^{-11}

表 4　主成分分析を用いた教師なし学習による変数選択でヒトに対して選択された 116 遺伝子を Enrichr にアップロードしたときの “GTEx Tissue Sample Gene Expression Profiles down” のカテゴリの上位5 位までの結果

Term	Overlap	P-value	Adjusted P-value
GTEX-Q2AG-0011-R10A-SM2HMLA_brain_female_40-49_years	51/1467	1.47×10^{-27}	3.29×10^{-24}
GTEX-TSE9-3026-SM3DB76_brain_female_60-69_years	49/1384	1.06×10^{-26}	1.19×10^{-23}
GTEX-S7SE-0011-R10A-SM2XCDF_brain_male_50-59_years	44/1278	3.20×10^{-23}	1.43×10^{-20}
GTEX-QMR6-1426-SM32PLA_brain_male_50-59_years	41/1066	2.57×10^{-23}	1.43×10^{-20}
GTEX-RNOR-2326-SM2TF4I_brain_female_50-59_years	47/1484	2.02×10^{-23}	1.43×10^{-20}

表 5　主成分分析を用いた教師なし学習による変数選択でマウスに対して選択された 118 遺伝子を Enrichr にアップロードしたときの “GTEx Tissue Sample Gene Expression Profiles down” のカテゴリの上位5 位までの結果

Term	Overlap	P-value	Adjusted P-value
GTEX-U8XE-0126-SM-4E3I3_testis_male_30-39_years	15/376	6.13×10^{-9}	3.45×10^{-6}
GTEX-X4XX-0011-R10B-SM46MWO_brain_male_60-69_years	23/938	5.25×10^{-9}	3.45×10^{-6}
GTEX-U4B1-1526-SM4DXSL_testis_male_40-49_years	13/282	1.23×10^{-8}	3.71×10^{-6}
GTEX-Q2AG-0011-R10A-SM2HMLA_brain_female_40-49_years	29/1467	5.11×10^{-9}	3.45×10^{-6}
GTEX-RNOR-2326-SM2TF4I_brain_female_50-59_years	29/1484	6.62×10^{-9}	3.45×10^{-6}

では２３個の転写因子が選ばれている。数には倍の差があるが、共通性は非常に高い。マウスで選ばれた２３個の転写因子のうち、KLF 以外の２２個は全てヒト側でも選ばれている。全く別の遺伝子発現プロファイルの解析から得られた結果であることを考えると、ヒトとマウスの共通の脳の発生機構を捉えている可能性は高い。

実際にこれらが制御ネットワークを構成しているかを見るために、主成分分析を用いた教師なし学習による変数選択でヒトに対して選択された 116 遺伝子とマウスに対して選択された 118 遺伝子を RegNetwork [7] にアップロードしてみた(図 **1**)。その結果、既知の制御関係を反映したネットワークがちゃんと検出されていることが判明した。

表 6　主成分分析を用いた教師なし学習による変数選択でヒトに対して選択された 116 遺伝子とマウスに対して選択された 118 遺伝子を Enrichr にアップロードしたときの ‘Jensen TISSUES” の Embryonic_brain の結果

Term	Overlap	*P*-value	Adjusted *P*-value
Human			
Embryonic_brain	71/4936	2.52×10^{-16}	4.07×10^{-15}
Mouse			
Embryonic_brain	75/4936	8.90×10^{-20}	1.06×10^{-18}

表 7　主成分分析を用いた教師なし学習による変数選択でヒトに対して選択された 116 遺伝子とマウスに対して選択された 118 遺伝子を Enrichr にアップロードしたときの “ENCODE and ChEA Consensus TFs from ChIP-X” の結果(adjusetd *P*-values が 0.01 以下の遺伝子)。太字はマウスとヒトで共通に選ばれたもの

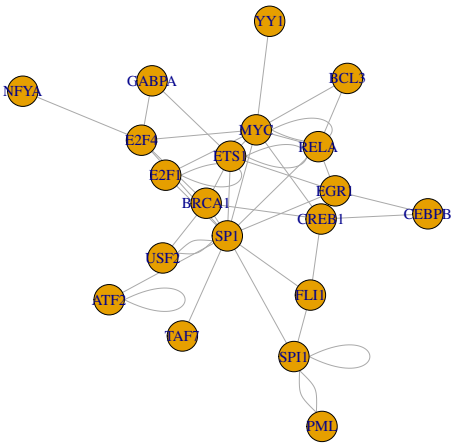
Human	**ATF2**, **BCL3**, BCLAF1, BHLHE40, **BRCA1**, **CEBPB**, **CEBPD**, **CHD1**, **CREB1**, CTCF, **E2F1**, E2F4, **EGR1**, ELF1, ETS1, FLI1, GABPA, **KAT2A**, KLF4, **MAX**, **MYC**, NANOG, **NELFE**, NFYA, NFYB, NR2C2, **PBX3**, **PML**, **RELA**, SALL4, SIN3A, SIX5, SOX2, SP1, SPI1, **TAF1**, **TAF7**, **TCF3**, USF2, **YY1**, ZBTB33, **ZMIZ1**
Mouse	**ATF2**, **BCL3**, **BRCA1**, **CEBPB**, **CEBPD**, **CHD1**, **CREB1**, **E2F1**, **EGR1**, **KAT2A**, KLF, **MAX**, **MYC**, **NELFE**, **PBX3**, **PML**, **RELA**, **SIN3A**, **TAF1**, **TAF7**, **TCF3**, **YY1**, **ZMIZ1**

最後に、検出された転写因子と胎児の脳の発生との関連が既報とどの程度合うのかみてみた。TAF7 は胎児の発生に重要な役割を垂らしていることが報告されている [8]。KAT2A, ATF2 と TAF1 は脳の発生に重要な役割があることが報告されている [9]。 BRCA1 も脳の発生に重要な役割があることが報告されている [10]。 一方、CEBPD や CREB は脳疾患に関係していることが報告されている [11,12]。E2F1 は出産後の脳の発生に関係している [13]。EGR1 もまた、脳での発現が報告されている [14]。PML と SIN3A も脳の発生との関連が報告され [15,16]、TCF3 はゼブラフィッシュの脳の発生過程への寄与が報告されている [17]。更に、YY1 は脳の発生への寄与が報告されている [18]。要するにヒトとマウスで共通に同定された転写因子はほとんど全部、脳の発生に関係しているという報告があるのである。このことから主成分分析を用いた教師なし学習による変数選択は一細胞 RNA-seq のデータから重要な遺伝子を選択するツールとして非常に優れていることがわかる。

4. 他の手法との比較

それでは他手法はこのデータセットに用いた場合、どのような性能をあげるだろうか。まずは highly variable genes を試した。その結果、ヒトに対しては１６８遺伝子、マウスに対しては１７１遺伝子が選ばれた。この数は主成分分析を用いた教師なし学習による変数選択で選ばれた数と同程度であり、妥当な数だと思われる。実際、共通に選ばれた遺伝子も４４遺伝子あった。共通に選ばれた遺伝子の割合は主成分分析を用いた教師なし学習による変数選択で共有に選ばれた割合よりは少ないものの、偶然に選ばれるにしては非常に多数であることには変わりはない。一方、主成分分析を用いた教師なし学習による変数選択で選ばれた遺伝子とはヒトとマウスでそれぞれ、たった4 遺伝子しか

ヒトの転写因子ネットワーク



マウスの転写因子ネットワーク

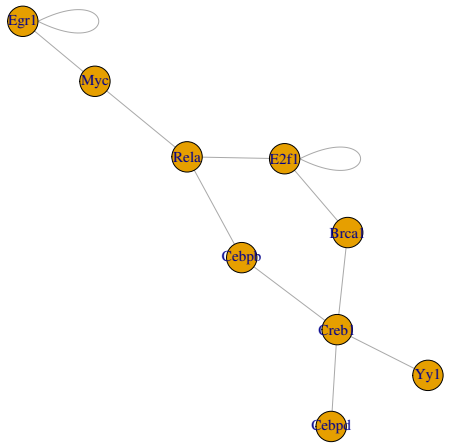


図 1　主成分分析を用いた教師なし学習による変数選択でヒトに対して選択された 116 遺伝子とマウスに対して選択された 118 遺伝子を RegNetwork にアップロードした時の制御関係ネットワーク。上: ヒト、下: マウス。Databases,(制御関係) Type,Evidence は All を選び、Organism はそれぞれヒトとマウスを指定し、Confidence は High を指定した。

被ってはいなかった。このことから、highly variable genes で選択された遺伝子については別途検討が必要と思われる。そこで、主成分分析を用いた教師なし学習による変数選択のときと同じように、選ばれた遺伝子の生物学的な妥当性を評価した。その結果はかなり劣ったものであった（詳しくは国際会議論文 [1,2] とその Supplementary material を参照のこと）。次に、Bimodal gene を試した。その結果は、ヒトに対して 11344 遺伝子、マウスに対して 10849 遺伝子と多すぎる遺伝子が選択されることが解った。このことから Bimodal gene は一細胞 RNA-seq から遺伝子を選択するにはあまりいい方法ではないことが伺われる。しかし、そこで生物学的な妥当性を検討するため、あえて上位２００遺伝子ずつを選んで評価を試みた。まず、共通に選ばれた遺伝子が２１遺伝子しかないことが解った。偶然に選ばれるにしては多すぎる数ではあるが、主成分分析を用いた教師なし学習による変数選択で共通に選ばれた５３遺伝子、highly variable genes で共通に選ばれた４４遺伝子にくらべると（ヒトとマウスでそれぞれ選ばれた遺伝子は多いにも関わらず）半分以下であり、選択の妥当性も低下していることが伺われる。実際、生物学的な評価はかなり劣っていることが解った（詳しくはこちらも、国際会議論文 [1,2] とその Supplementary material を参照のこと）。最後に dpFeature を試した。これは一細胞 RNA-seq のために考えられた方法であり、highly variable genes や bimodal genes より性能が高いことが期待される。しかし、実際にやってみるとヒトに対しては１３７７５遺伝子、マウスに対しては１３３６２遺伝子が選ばれてしまい、十分な遺伝子数の絞り込みにはつかえないことがわかった。それでも生物学的な妥当性を評価するため、Bimodal genes の時のように、ヒトとマウスぞれぞれ、上位２００遺伝子ずつを選んでみた。共通に選ばれた遺伝子数は７６遺伝子であり、主成分分析を用いた教師なし学習による変数選択や highly variable genes に比べて遜色ない高い割合であった。しかし、残念ながら実際に生物学的な評価を行うとその結果は芳しいものではなかった（詳しくはこちらも、国際会議論文 [1,2] とその Supplementary material を参照のこと）。

5. おわりに

本稿では主成分分析を用いた教師なし学習による変数選択を一細胞 RNA-seq に適用した。その結果、生物学的な妥当性の高い、ごく限られた数の遺伝子を選択できることが解った。また、同手法を highlt variable genes や bimodal genes、dpFeature と比較したが、生物学的な妥当性は主成分分析を用いた教師なし学習による変数選択が一番であった。実際、国際会議論文 [1,2] 刊行後、主成分分析を用いた教師なし学習による変数選択は、別の研究者によって、一細胞 RNA − seq に対する性能比較を他手法と行われ、同等程度の性能があると評価されたことを付記する [19]。

謝辞　本研究の原著論文は科研費基盤研究 (C) 17K00417 の助成を受けて行われた。

参考文献

[1] Taguchi, Y.-h.: Principal Component Analysis-Based Unsupervised Feature Extraction Applied to Single-Cell Gene Expression Analysis, *Intelligent Computing Theories and Application* (Huang, D.-S., Jo, K.-H. and Zhang, X.-L., eds.), Cham, Springer International Publishing, pp. 816–826 (2018).

[2] Taguchi, Y.-h.: Principal component analysis-based unsupervised feature extraction applied to single-cell gene expression analysis, *bioRxiv*, (online), DOI: 10.1101/312892 (2018).

[3] Manno, G. L., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L. E., Stott, S. R., Toledo, E. M., Villaescusa, J. C., Lnnerberg, P., Ryge, J., Barker, R. A., Arenas, E. and Linnarsson, S.: Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells, *Cell*, Vol. 167, No. 2, pp. 566–580.e19 (online), DOI: 10.1016/j.cell.2016.09.027 (2016).

[4] 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).

[5] Chen, H.-I. H., Jin, Y., Huang, Y. and Chen, Y.: Detection of high variability in gene expression from single-cell RNA-seq profiling, *BMC Genomics*, Vol. 17, No. 7, p. 508 (online), DOI: 10.1186/s12864-016-2897-6 (2016).

[6] Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. and Trapnell, C.: Reversed graph embedding resolves complex single-cell trajectories, *Nature Methods*, Vol. 14, No. 10, pp. 979–982 (online), DOI: 10.1038/nmeth.4402 (2017).

[7] Liu, Z.-P., Wu, C., Miao, H. and Wu, H.: RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse, *Database*, Vol. 2015 (online), DOI: 10.1093/database/bav095 (2015).

[8] Gegonne, A., Tai, X., Zhang, J., Wu, G., Zhu, J., Yoshimoto, A., Hanson, J., Cultraro, C., Chen, Q.-R., Guinter, T., Yang, Z., Hathcock, K., Singer, A., Rodriguez-Canales, J., Tessarollo, L., Mackem, S., Meerzaman, D., Buetow, K. and Singer, D. S.: The General Transcription Factor TAF7 Is Essential for Embryonic Development but Not Essential for the Survival or Differentiation of Mature T Cells, *Molecular and Cellular Biology*, Vol. 32, No. 10, pp. 1984–1997 (online), DOI: 10.1128/MCB.06305-11 (2012).

[9] Tapias, A. and Wang, Z.-Q.: Lysine Acetylation and Deacetylation in Brain Development and Neuropathies, *Genomics, Proteomics & Bioinformatics*, Vol. 15, No. 1, pp. 19 – 36 (online), DOI: https://doi.org/10.1016/j.gpb.2016.09.002 (2017).

[10] Pao, G. M., Zhu, Q., Perez-Garcia, C. G., Chou, S.-J., Suh, H., Gage, F. H., O'Leary, D. D. M. and Verma, I. M.: Role of BRCA1 in brain development, *Proceedings of the National Academy of Sciences*, Vol. 111, No. 13, pp. E1240–E1248 (online), DOI: 10.1073/pnas.1400783111 (2014).

[11] Sun, Y., Jia, L., Williams, M. T., Zamzow, M., Ran, H.,

Quinn, B., Aronow, B. J., Vorhees, C. V., Witte, D. P. and Grabowski, G. A.: Temporal gene expression profiling reveals CEBPD as a candidate regulator of brain disease in prosaposin deficient mice, *BMC Neuroscience*, Vol. 9, No. 1, p. 76 (online), DOI: 10.1186/1471-2202-9-76 (2008).

[12] Mantamadiotis, T., Lemberger, T., Bleckmann, S. C., Kern, H., Kretz, O., Villalba, A. M., Tronche, F., Kellendonk, C., Gau, D., Kapfhammer, J., Otto, C., Schmid, W. and Schtz, G.: Disruption of CREB function in brain leads to neurodegeneration, *Nature Genetics*, Vol. 31, No. 1, pp. 47–54 (online), DOI: 10.1038/ng882 (2002).

[13] Suzuki, D. E., Ariza, C. B., Porcionatto, M. A. and Okamoto, O. K.: Upregulation of E2F1 in cerebellar neuroprogenitor cells and cell cycle arrest during postnatal brain development, *In Vitro Cellular & Developmental Biology - Animal*, Vol. 47, No. 7, pp. 492–499 (online), DOI: 10.1007/s11626-011-9426-3 (2011).

[14] Wells, T., Rough, K. and Carter, D.: Transcription Mapping of Embryonic Rat Brain Reveals EGR-1 Induction in SOX2+ Neural Progenitor Cells, *Frontiers in Molecular Neuroscience*, Vol. 4, p. 6 (online), DOI: 10.3389/fnmol.2011.00006 (2011).

[15] Korb, E. and Finkbeiner, S.: PML in the Brain: From Development to Degeneration, *Frontiers in Oncology*, Vol. 3, p. 242 (online), DOI: 10.3389/fonc.2013.00242 (2013).

[16] Witteveen, J. S., Willemsen, M. H., Dombroski, T. C. D., van Bakel, N. H. M., Nillesen, W. M., van Hulten, J. A., Jansen, E. J. R., Verkaik, D., Veenstra-Knol, H. E., van Ravenswaaij-Arts, C. M. A., Wassink-Ruiter, J. S. K., Vincent, M., David, A., Caignec, C. L., Schieving, J., Gilissen, C., Foulds, N., Rump, P., Strom, T., Cremer, K., Zink, A. M., Engels, H., de Munnik, S. A., Visser, J. E., Brunner, H. G., Martens, G. J. M., Pfundt, R., Kleefstra, T. and Kolk, S. M.: Haploinsufficiency of MeCP2-interacting transcriptional co-repressor SIN3A causes mild intellectual disability by affecting the development of cortical integrity, *Nature Genetics*, Vol. 48, No. 8, pp. 877–887 (online), DOI: 10.1038/ng.3619 (2016).

[17] Dorsky, R. I., Itoh, M., Moon, R. T. and Chitnis, A.: Two tcf3 genes cooperate to pattern the zebrafish brain, *Development*, Vol. 130, No. 9, pp. 1937–1947 (online), DOI: 10.1242/dev.00402 (2003).

[18] Beagan, J. A., Duong, M. T., Titus, K. R., Zhou, L., Cao, Z., Ma, J., Lachanski, C. V., Gillis, D. R. and Phillips-Cremins, J. E.: YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment, *Genome Research*, Vol. 27, No. 7, pp. 1139–1152 (online), DOI: 10.1101/gr.215160.116 (2017).

[19] Chen, B., Lau, K. S. and Herring, C. A.: pyNVR: investigating factors affecting feature selection from scRNA-seq data for lineage reconstruction, (online), DOI: 10.1093/bioinformatics/bty950 (2018).