Vol.2021-MPS-133 No.30
Vol.2021-BIO-66 No.30
2021/6/30

PCAUFEを用いたCOVID-19と他の肺疾患を区別する遺伝子の特定

志茂　衛[1,a)]　藤澤　孝太[2]　田口　善弘[3]　池松　真也[4]　宮田　龍太[5]

概要：本研究ではCOVID-19特有の疾患関連遺伝子を選択するために, COVID-19患者とそうでない肺疾患患者のデータセットにprincipal-components-analysis-based unsupervised feature extraction (PCAUFE)を適用した．19,472個の候補に対してサンプル数が126人と圧倒的に少ないデータセットでも145個がCOVID-19関連遺伝子として選択できた．またクラスタリング解析により, PCAUFEで選択された遺伝子のみで患者と非患者に線形分離可能であることも確認した．Enrichrを用いたエンリッチメント解析によりこれら145個の遺伝子の調節転写因子を調べたところSPI1, CEBPA, STAT1, STAT2などの免疫系が上位にあった．

Gene selection to distinguish COVID-19 from other lung diseases using PCAUFE

1. はじめに

現在, 新型コロナウイルス感染症（coronavirus disease 2019,COVID-19）が世界中で猛威を振るっており, 特効薬の開発が急務となっている．創薬を行うにあたり疾患関連遺伝子の特定が必須である．現在COVID-19で報告されているデータは患者のサンプル数が候補となる遺伝子数より圧倒的に少ない．そのため, サンプルを大量に要する従来の教師あり学習では有意な遺伝子選択は困難である．

そこで先行研究ではFujisawaらは, 主成分分析を基にした教師無し学習法principal-components-analysis-based unsupervised feature extraction (PCAUFE)[1]に, COVID-19患者と健常者のデータセットを適用して遺伝子選択を行った．その結果, サンプル数が34人と少ない場合でも60,683個の候補となる遺伝群から123個のCOVID-19疾患関連遺伝子を選択できた[2].

しかしこの123個の遺伝子群では, COVID-19感染者と何の病気にも感染していない健常者を分けることはできるが, 別の肺の病気を患っている人を厳密に区別できない．

そこで本研究では健常者のデータセットを使用した場合よりさらにCOVID-19特有の疾患関連遺伝子を選択するために, COVID-19患者とそうでない肺疾患患者のデータセットをPCAUFEに適用して遺伝子選択を行った．さらに, それらを上流で制御している転写因子も調べた．またPCAUFEで選択された遺伝子を別のCOVID-19のデータセットに適用しCOVID-19患者とCOVID-19非患者とで分類できるか試した．

2. データと方法

2.1 PCAUFE[1]

PCAUFEは, サンプルよりも変数が多い（$N << P$）状況の下, 後述する工夫をPCAに施すことで, 患者と非患者といった2群間で有意差が確認できる変数の選択を目的とする（詳細は[1]を参照）．まず行を変数, 列をサンプルとしたデータセット$\boldsymbol{x}$の共分散行列の第j $(= 1, 2, \cdots, N)$固有値および固有ベクトル（主成分負荷量）$\boldsymbol{w}_j$を計算する．次に$\boldsymbol{w}_j$をあらかじめ各サンプルに付与したラベルに基づいて2群に分類し,t検定で有意差が確認できる序数jを探す．そして, 各変数i $(= 1, 2, \cdots, P)$の第j主成分スコア$(t_j = \boldsymbol{x}_i \cdot \boldsymbol{w}_j)$に対してそれぞれカイ二乗検定を行い$P$値を算出し, 有意水準未満の値になった変数のみを選択する．

[1] 琉球大学大学院理工学研究科
[2] 東京工業大学大学院生命理工学院
[3] 中央大学理工学部
[4] 沖縄工業高等専門学校生物資源工学科
[5] 琉球大学工学部
[a)] k218439@eve.u-ryukyu.ac.jp

Vol.2021-MPS-133 No.30
Vol.2021-BIO-66 No.30
2021/6/30

2.2 データセット

本研究で用いた二つの mRNA プロファイルはいずれも被験者の血液から採取されたもので, NCBI GEO [3] から取得した．片方は PCAUFE で COVID-19 に関連する遺伝子群を選択するために, もう片方はその選択遺伝子群の COVID-19 患者 · 非患者識別能力をクラスター分析で評価するために使用した．

クラスター分析用プロファイル GSE157103[4] は COVID-19 の患者 100 名と COVID-19 ではない肺疾患患者 26 名からなり，遺伝子候補数は 19,472 個であった．PCAUFE を適用する前処理として, データセットに標準化を施した．

患者 · 非患者分類解析用プロファイル GSE152418[5] は原ラベル Healthy, Convalescent, Moderate, Severe, ICU のうち, 前者二つを COVID-19 患者（17 名）とし, 残りを非患者（17 名）とした．

3. 結果

PCAUFE をデータセット GSE157103[4] に適用したところ，19,472 個の候補から 145 個が COVID-19 と他の肺疾患を区別する遺伝子として選択された．これらの遺伝子について GeneSet DB[6] を用いてエンリッチメント解析を行ったところ，免疫系の GO term を有するものが多かった．

次にデータセット GSE152418 において，選択遺伝子 145 個のみで患者 · 非患者分類を行った結果を図 1 に示す．この図より，選択遺伝子 145 個のみで COVID-19 患者 · 非患者について線形分離可能であることが確認できる．

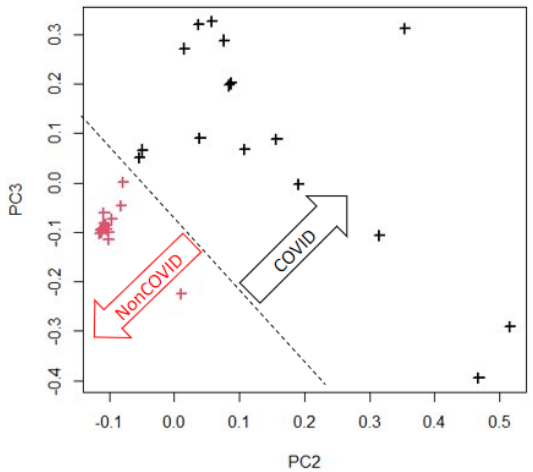

図 1 PCAUFE で選択された 145 個の遺伝子のみで GSE152418[5] のデータセットをクラスター分析した結果． 点線部分で COVID-19 患者と非患者を完全分離できることを確認した．

4. 考察

PCAUFE で選択された 145 個の遺伝子について Enrichr[7] の TRRUST Transcription Factors 2019 と ENCODE TF ChiP-seq2015 を用いて転写因子を調べた結果, SPI1, CEBPA, STAT1, STAT2 が上位にあった．

SPI1 は骨髄系および B リンパ系細胞の発達中に遺伝子発現を活性化する ETS ドメイン転写因子をコードしており, CEBPA は増殖停止と骨髄系前駆細胞, 脂肪細胞, 肝細胞, 肺と胎盤の細胞の分化を調整する転写因子である．SPI1 と CEBPA はともに骨髄前駆細胞に存在しており, 文献 [8] では CEBPA は SPI1 と物理的に相互作用し, SPI1 の転写活性を低下させ SPI1 誘発樹状細胞の発達をブロックすることが知られている．また COVID-19 の原因である SARS-CoV-2 の急性感染は単球, 樹状細胞の比率を機能障害に伴って減少させ, T 細胞, ナチュラルキラー細胞を含む広範な免疫細胞の減少させることが文献 [9] で報告されている．

STAT2 は I 型 IFN（IFN- α および IFN- β）によるシグナル伝達を媒介するシグナル伝達物質および転写活性化因子であり COVID-19 において, ハムスターとマウスの実験では STAT2 シグナル伝達が一方では SARS-CoV-2 誘発性肺疾患を促進させ, もう一方では全身性ウイルスの伝播を制限しているという二重の役割を果たしていることが分かっている [10]．また STAT1 は I, II 型, または III 型インターフェロンのいずれかによるシグナルによる遺伝子のアップレギュレーションに関与しており, COVID-19 では SARS-CoV-2 遺伝子産物である NSP1 および ORF6 タンパク質などのよって, 機能機能不全を誘発される [11]．

まとめとして, 本研究では PCAUFE を用いて COVID-19 と他の肺疾患とを区別する 145 個の遺伝子をデータドリブンに特定し, エンリッチメント解析によりそれらの調節転写因子として SPI1, CEBPA, STAT1, STAT2 といった免疫系が上位にあったことを特定した．

参考文献

[1] Taguchi Y (2017) *Scientific Reports*. 7:44016-44029.
[2] Fujisawa K et al. (2021) under review.
[3] Gene Expression Omnibus, NCBI (online), available from https://www.ncbi.nlm.nih.gov/geo
[4] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157103
[5] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152418
[6] Araki H, Knapp C, Tsai P, Print C (2012) *FEBS Open Bio*. 2:76-82
[7] Edward Y Chen, Christopher M Tan et al. (2013)
[8] Venkateshwar A, Reddy, Iwama A, Iotzova G et al. (2002) *Blood*.100 (2) : 483–490.
[9] Runhong Zhou, Kelvin Kai-Wang To, Yik-Chun Wong et al. (2020)*immunity*. 4:864-877.
[10] Robbert B et al. (2020) *Nature Communications*. 11:5838.
[11] Toshifumi M et al. (2020) *Cell Death Differentiation*. 27:3209–3225.