［課程－２］

審査の結果の要旨

氏名 波多野　裕明

本研究は, 全身性強皮症, 全身性エリテマトーデス, 炎症性筋炎など 10 種の全身性自己免疫疾患の患者および健常者, 計 418 人の末梢血から, 27 種の免疫担当細胞サブセットを分取し, そのトランスクリプトームについて機械学習を用いて解析したものであり, 下記の結果を得ている.

１．解析対象とした 9,897 サンプルについて, 遺伝子発現による階層型クラスタリングを行ったところ, 各サンプルはそれぞれの親サブセット (CD4 陽性 T 細胞, CD8 陽性 T 細胞, B 細胞, NK 細胞, 好中球, 単球, 樹状細胞)と対応した群にクラスタリングされた. さらに, 他サブセットを自身の部分集合として含まない 24 サブセット 8,923 サンプルについて階層型クラスタリングを行ったところ, 各サブセットがクラスターを形成し, ほぼ全サンプルがサブセットをまたぐことなくクラスタリングされた. これらの結果から, 24 の異なるサブセットがそれぞれ特徴的な遺伝子発現を有すること, ほぼ全てのサンプルが正確に各サブセットの遺伝子発現の特徴を捉えていることが確認された.

２．64 例の健常者のサンプルを用いて, サブセット特異的な発現パターンをとる遺伝子を検出した. 機械学習の１つであるランダムフォレスト (RF)により, あるサブセットを他のサブセットと弁別する試行を 1,000 回行ない, 遺伝子を弁別重要度によってランキングすることにより, 各サブセットを特徴付けるのに重要な上位 100 遺伝子を同定した. 同定した遺伝子リストには, 既知のマスターレギュレーター遺伝子が上位に含まれており, 結果の妥当性が確認された. また, long non-coding RNA (lncRNA)も含まれており, サブセット特異的な発現パターンをとる lncRNA が存在することが確認された.

３．疾患の違いが遺伝子発現に与える影響について検討した. サンプルが 300 以上存在する 19 サブセットを対象に, 各サブセットで主成分分析を行い, 上位 20 個の主成分を算出した. 計 380 個の各主成分について, 疾患を説明変数として主成分得点を線形回帰した結果, 140 個の主成分がいずれかの疾患の有無と有意に (FDR<0.01)関連した. この結果から, 疾患による遺伝子発現の変動が, 本データセットの遺伝子発現の分散に大きく寄与していることが分かった.

４．同定した 140 個の主成分を用いて, 疾患および個人の階層型クラスタリングを行うことで, 疾患間の類似性や疾患内の多様性について検討した. 疾患のクラスタリングでは, 全身性エリテマトーデス (SLE)と混合性結合組織病 (MCTD)の類似性が高く, ついで炎症性筋炎 (Myo), 全身性強皮症 (SSc), シェーグレン症候群 (SjS), 関節リウマ

チ (RA)といった自己免疫疾患が隣接し，一方でこれらと離れて成人スティル病 (AOSD)とベーチェット病 (BD), ANCA 関連血管炎 (AAV)と高安動脈炎 (TAK)という自己炎症性疾患および血管炎がクラスタリングされた．これは，臨床的な疾患分類と合致する結果であった．さらに，各主成分がどのような遺伝子群の発現を反映しているかについて概観するため，因子負荷量に応じたエンリッチメント解析を行った．結果として, SLE と MCTD に共に関連する主成分がインターフェロン関連の遺伝子群の発現を反映していることや, BD, AAV, AOSD に共に関連する主成分の一部が, Cell Cycle や Translation activity 関連の遺伝子群の発現を反映していることが示唆された．個人のクラスタリングでは，同じ疾患では比較的隣接した位置にクラスタリングされる傾向があったが，疾患ごとに１つにはまとまらず，個人間の分散が大きいことが推察され，患者間の病態の多様性が反映された結果と考えられた．

４．50 症例以上のデータがある 3 疾患 (SLE, SSc, Myo)を対象として，遺伝子発現及びスプライシングイベントを入力に用いた RF により, i) 健常者 (HC) および ii) その他の疾患 (others)との間の弁別を行った．対象疾患群をトレーニングデータ用，テストデータ用に 6：4 にランダムに分割し，対象疾患群とサンプルサイズ・男女比が一致するように比較対象群もランダム抽出した．このようなランダム抽出を 100 回施行し，各回でトレーニングデータによる予測モデルの作成と重要度算出を行い，テストデータによる精度評価を行った．結果として, SLE と HC では 95%, Myo と HC では 90%, SSc と HC では 85%の予測精度が得られた．これらの疾患は治療中であったとしても，免疫細胞の遺伝子発現が健常者とは大きく異なっていることが示唆された.また, SLE と others では 90%, Myo と others, SSc と others でも 70%前後の予測精度が得られ，各疾患に特異的な特徴量が存在することが示された．

５．SLE, Myo, SSc の弁別に用いた(サブセット)__(遺伝子)または(サブセット)__(スプライシングイベント)の組を 100 回の試行での重要度の平均によりランキングすることにより，これらの疾患に特徴的な遺伝子発現，スプライシングイベントを網羅的に同定し，いくつかの遺伝子において，実際に各疾患での発現に有意差があることを示した．また，疾患弁別重要度が上位であったスプライシングイベントについて，ゲノムにマップされた全リードを疾患別に重ねて可視化することにより, SLE, Myo にそれぞれ特徴的なスプライシングイベントが存在することを示した．

以上，本研究は, 10 種の自己免疫疾患を含む 27 種の末梢血免疫細胞サブセットのトランスクリプトームデータという，過去に報告がない大規模なデータセットを作成し，疾患横断的な解析を行った.トランスクリプトームデータにより高精度に疾患を弁別することが可能であることを示し, SLE, Myo, SSc における疾患特異的な遺伝子発現やスプライシングイベントを網羅的に同定し，各疾患の病態の解明につながる可能性のある新規知見を示した．
よって本論文は博士 (医学)の学位請求論文として合格と認められる．