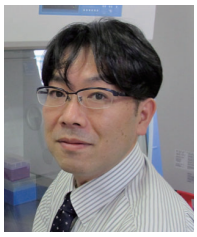
［研究助成（B）］

# エピゲノム情報を基にした機械学習によるヒトiPS細胞の肝細胞分化効率予測システム構築のための基礎研究

Study of a prediction system for hepatocyte differenciation efficiency of human iPSCs by machine learning based on epigenome information

2191901



研究代表者　宮崎大学　農学部　教　授　西　野　光一郎

## ［研究の目的］

本研究は，ヒト人工多能性幹細胞（induced pluripotent stem cell：iPS細胞）のロット間における特性の違いを機械学習により識別し，肝細胞分化効率を予測するシステム構築のための基礎研究である。ヒトiPS細胞は再生医療における切り札であるが，樹立されたロット間には明確な特性の違いが存在する。その特性は，細胞分化効率の差として顕著に現れ，ヒトiPS細胞の実用化の大きな壁となっている。本研究では，細胞内バイオデータと機械学習技術を融合することで，未分化状態のiPS細胞から肝細胞への分化効率を予測するシステムを開発する。多数の未分化ヒトiPS細胞から網羅的DNAメチル化情報と，肝細胞への分化効率を取得し，教師あり学習を行うことで，精度の高い学習モデルの構築を行う。

## ［研究の内容，成果］

肝疾患による年間死亡数は17万人を超えており，死因では常に10位以内に入る（令和元年度厚生労働省「人口動態統計の概況」）。肝疾患の発生メカニズムを詳細に解明することは肝疾患の克服に向けての大きな課題である。肝細胞を *in vitro* で再現し，そのメカニズムを解明するツールとしてiPS細胞は有用である。

しかし，同一の親細胞を使用し，同一の作成方法で樹立し，同一の培養条件で得られたヒトiPS細胞であっても，細胞ロット間によって分化能力に違いが存在する。再生医療を目的にヒトiPS細胞から様々な細胞への分化誘導系が研究開発されているが，ヒトiPS細胞ロット間の分化能力の違い（分化指向性）は，再生医療を実用化する上で大きな障害となっている（図1）。

このヒトiPS細胞の肝細胞分化誘導において分化指向性が存在するのかを検討した。十分に細胞数を確保したヒトiPS細胞をアクチビンA，Wnt3a，HGF等を添加したRPMI/B27培地で培養し，内胚葉へ分化誘導し，その後，オンコスタチンやデキサメタゾン等を含む2種類

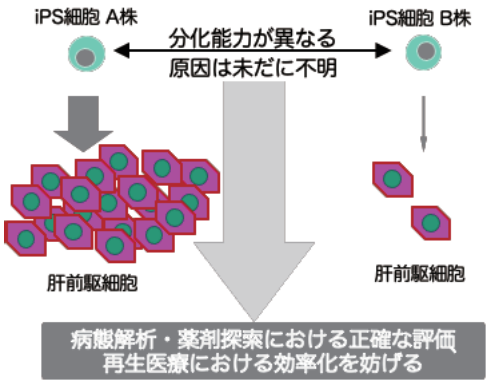



図1　iPS細胞株間の分化指向性がにもたらす問題

の肝細胞分化培地で培養することで肝細胞を誘導した。

分化誘導培養後，細胞を固定し，肝幹細胞のマーカーである AFP に対する抗 AFP 抗体で免疫化学染色を施した。フローサイトメーターを用いて染色細胞を解析し，分化効率（AFP 陽性細胞の割合）を測定した。研究期間を通して合計 18 株のヒト iPS 細胞について肝細胞分化誘導実験を行い，検証したところ，その分化誘導効率に大きな違いが存在することが明らかとなった。18 株の内，明確な低分化効率株 6 株および高分化効率株 6 株の合計 12 株を以降の実験に用いた（図 2）。

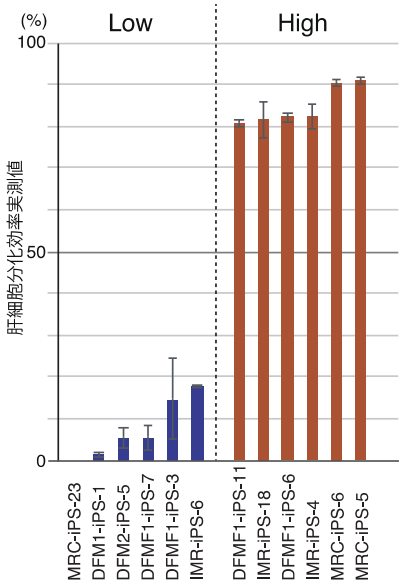



図 2　iPS 細胞株間の肝細胞分化効率の比較

上記のヒト iPS 細胞 12 株について，大量に増やした未分化ヒト iPS 細胞を回収し，ゲノム DNA を抽出後，バイサルファイト反応を施し，Illumina Infinium MethylationEPIC BeadChip を用いて DNA メチル化プロファイルを得た。Infinium MethylationEPIC BeadChip は，ヒトゲノム DNA 上の 85 万箇所以上の CpG 部位のメチル化率を測定できるアレイシステムである。得られた網羅的 DNA メチル化データを用いて階層的クラスタリング解析を行った。階層的クラスタリング解析の結果と図 1 の肝細胞分化効率の結果を合わせた結果が図 3 である。

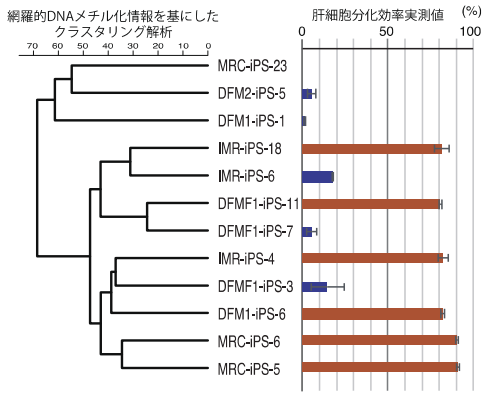



図 3　iPS 細胞株間の DNA メチル化クラスタリング解析と肝細胞分化効率

図 3 の結果で明らかなように教師なし学習である階層的クラスタリング解析による分類では，ヒト iPS 細胞の肝細胞分化効率を予測することは不可能である。また，未分化ヒト iPS 細胞における肝細胞への分化指向性を制御する因子は同定されていない。これらの結果は，ヒト iPS 細胞の分化指向性は単一の遺伝子発現に依存する単純なものではなく，細胞内の分子間ネットワーク，つまり遺伝子発現やエピジェネティクス制御のネットワークにより規定されていることを示唆している。よって，iPS 細胞ロット間の肝細胞分化における分化指向性を解析し，肝細胞分化を制御するには，細胞から得られる網羅的データを包括的に解析，検定，評価する事が必要となる。

ヒト iPS 細胞の肝細胞への分化指向性を包括的に正確に理解するためには，肝細胞への分化効率の実測値の収集，様々な網羅的データの蓄積とそれらを総合的に解析する方法が必要である。この解析を実現する手段として，人工知能（AI）技術が有効である。AI 技術とバイオビックデータを活用することで，株間の差を分子レベルで理解し，且つ分化効率に影響を及ぼす分子ネットワークの解析が可能であり，それらを応用して株間の分化指向性を可視化することで肝分化のメカニズムをより詳細に解明できる。

そこで本研究では，細胞内バイオデータと機械学習技術を融合することで，未分化状態の

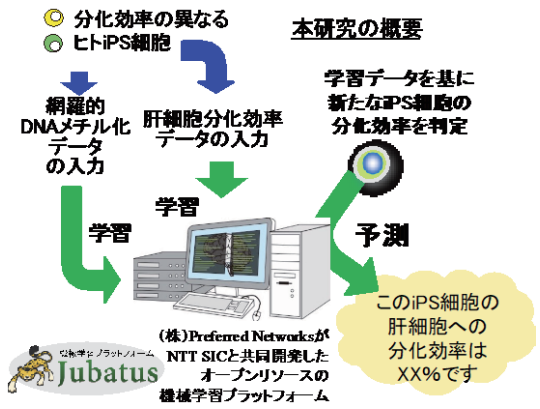



図 4　網羅的バイオデータと機械学習の融合による細胞特性判別システムの創出

iPS 細胞から肝細胞への分化効率を予測するシステムの開発を試みた。未分化ヒト iPS 細胞から網羅的 DNA メチル化情報と肝細胞への分化効率実測値を取得し，教師あり学習を行うことで，学習モデルの構築を行った（図 4）。

本研究における成否は，学習モデルを構築するための良質な学習データをいかに多く確保できるかにかかっている。本件で定義する良質な学習データとは，すなわち質を保証された iPS 細胞から得られた DNA メチル化データであり，同一 iPS 細胞株を分化誘導して得られる分化効率の実測値データである。培養環境や継代数でその性質が変動するヒト iPS 細胞においては，良質な学習データを得るためには，ヒト iPS 細胞の培養，回収，分化誘導実験を同一条件にして行わなければならない。それらを踏まえ，良質な学習データを得るためのヒト iPS 細胞培養に多くの時間を費やした。十分なゲノム DNA を得るために，ヒト iPS 細胞株 12 株をそれぞれ培養，増殖し，十分な細胞数を得ることができた。ヒト iPS 細胞の培養は，マウス胎児由来線維芽細胞（MEF）をフィーダー細胞として，Knockout-DMEM, 代替血清である Knockout-Serum Replacement（KSR）と繊維芽細胞増殖因子 FGF-2（basic FGF）を添加した Knockout-DMEM をベースとした培地を使用した。

大量に増やした未分化ヒト iPS 細胞の一部はゲノム DNA 抽出に用い，残りの未分化ヒト iPS 細胞は肝細胞分化実験に供した。抽出したゲノム DNA は，バイサルファイト反応を施し，Illumina Infinium MethylationEPIC BeadChip を用いて DNA メチル化プロファイルを得た。1 検体につき 85 万箇所以上の CpG 部位のメチル化率を取得した。同時に DNA メチル化情報を取得した同一 iPS 細胞株を用いて肝細胞への分化誘導実験を実施し，肝分化効率の実測値を測定した。具体的には，肝幹細胞への分化誘導を行い，肝幹細胞のマーカーである AFP の蛍光抗体を用いて免疫細胞染色を行った。

前述と同様に十分に細胞数を確保したヒト iPS 細胞をアクチビン A，Wnt3a，HGF 等を添加した RPMI/B27 培地で培養し，内胚葉へ分化誘導し，その後，オンコスタチンやデキサメタゾン等を含む 2 種類の肝細胞分化培地で培養することで肝細胞を誘導した。分化誘導培養後，細胞を固定し，抗 AFP 抗体で免疫化学染色を施した。フローサイトメーターを用いて染色細胞を解析し，分化効率を測定した。これらの実験によって，ヒト iPS 細胞 12 株について，各未分化 iPS 細胞の DNA メチル化情報と，それに対応する肝細胞分化効率のデータの取得を行った。複数のヒト iPS 細胞から，同一培養条件での分化誘導実験によって一貫した分化効率評価データとそれに完全対応する網羅的 DNA メチル化データのセットはほとんど報告がなく，非常に貴重なものである。ヒト iPS 細胞から分化誘導実験までを一貫して行うことで，細胞ロット間の性質の差異を正確に評価し，解析することが可能なデータセットを取得したことは，本研究の特色であり，利点である。

網羅的 DNA メチル化情報を入力データとして，機械学習技術を用いるための基盤設定，各種ハイパーパラメータの設定，データマイニング，学習手順の設定など最適な学習手順の構築及び学習モデルのバリデーション法の検討を行った。生命科学では検体数の制限があり，数万の iPS 細胞ロットとその分子データを揃え，

分化誘導実験を行うことは現実的には不可能である。

そこで，本研究ではニューラルネットワークと比較して少ないサンプル数にも対応可能な線形分類器による機械学習を検討した。未分化ヒト iPS 細胞の DNA メチル化データ（入力データ）とそれぞれの株に対応する肝細胞分化効率の実測値データ（教師データ）のセットを用いて，学習モデルの構築を行った。線形分類器の Platform はオンライン機械学習向け分散処理フレームワーク：Jubatus（http : //jubat.us/ja/）を用い，アルゴリズム AROW による多値分類と二値分類を検討した。教師データのラベルの組み合わせと Regularization weight および epoch 数の組み合わせによる学習モデルを 16,000 ほど作成し，学習モデルを評価した。学習モデルの評価は，Leave-one-out cross-validation を行い，F-score の比較検証により行った。その結果，肝細胞分化効率 20% 以下を低分化群，80% 以上を高分化群の 2 ラベルとして二値分類の学習を行った群の成績が良く，その内 F-score の高いもの上位 3 つの条件を図 5 に示す。Regularization weight : 1.0，200-epoch 設定の学習モデルが最も評価値が高いものとなった。

| regularization weight | 1.0 | 0.9 | 1.1 |
|---|---|---|---|
| epoch | 220 | 190 | 180 |
| F-Score | | | |
| Test sample | 0.727 | 0.677 | 0.665 |
| Learning sample | 0.947 | 0.925 | 0.897 |
| Accuracy | | | |
| Test sample | 0.667 | 0.667 | 0.583 |
| Learning sample | 0.947 | 0.924 | 0.894 |

図 5　学習モデルの評価

## ［今後の研究の方向，課題］

本研究では，12 サンプルを使用し，1 サンプルにつき 85 万箇所の DNA メチル化データ（85 万の特徴量数）を用いて学習モデルの構築を試みた。近年の機械学習によるビッグデータ解析シーンでは，特徴量数に対してその 100 倍以上のサンプル数を用いることが一般的であり，本研究のように特徴量数が圧倒的に多い場合には，過学習による問題があり，機械学習は不向きとされる。

しかし，画像データとは異なり細胞内分子データを用いた機械学習の応用を考えた場合，特にヒト iPS 細胞の様な特殊なサンプルにおいては，サンプル数よりデータの特徴量数が多くならざるを得ない。近年，次世代シークエンサー解析やマイクロアレイ技術が一般化し，生命科学研究者は網羅的なバイオデータを取得できるようになった。これら膨大なバイオデータ（1 サンプルに対して膨大な特徴量を含む）の解析に機械学習技術を応用することで，これまで明らかにできなかった生命現象を解き明かすことが期待でき，それ故，網羅的バイオデータの機械学習手法の開発は必要である。本研究では，12 サンプルを使用し，85 万箇所の DNA メチル化データを用いて学習モデルを作成し，肝細胞分化予測モデルの基盤技術を構築するができた。サンプル数がまだ少ないためモデルの精度は十分ではないが，今後，より精度の高い機械学習法の検討を行い，機械学習技術の生命科学分野への積極的な応用を進めていく。精度の高い肝細胞分化予測モデル構築に向けては，ヒト iPS 細胞サンプルデータの追加は必須であり，追加学習を進め，学習モデルから抽出した要素解析を通して肝分化効率に影響を及ぼすゲノム領域の抽出を行い，肝細胞分化メカニズムの解明へ繋げていきたい。

## ［成果の発表，論文等］

［1］ Nishino K, Takasawa K, Okamura K, Arai Y, Sekiya A, Akutsu H, Umezawa A. Identification of an epigenetic signature in human induced pluripotent stem cells using a linear machine learning model. *Human Cell*. 34(1) : 99-110. 2021.