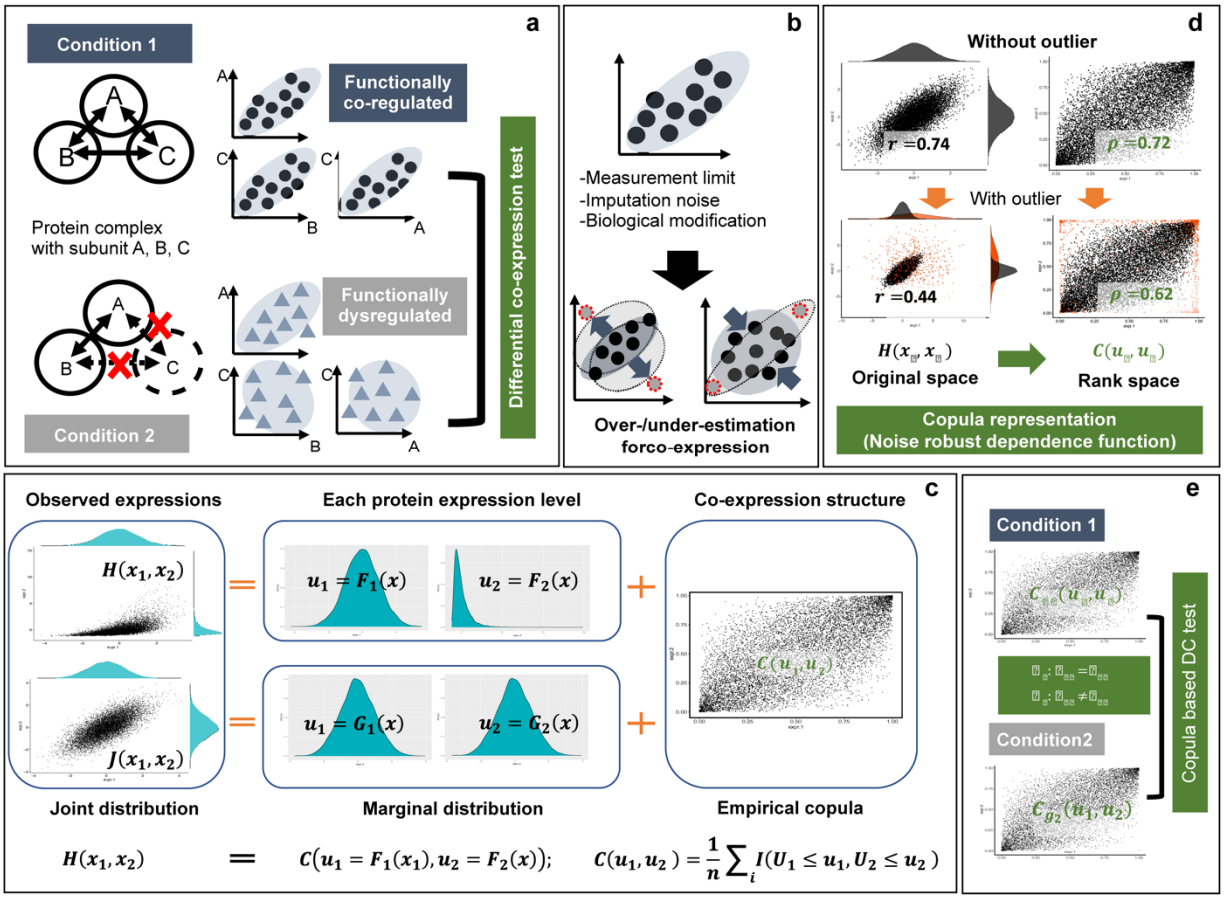

がん治療における標的タンパク質予測 AI の開発に成功！
～ポストゲノム時代の大規模がんプロテオミクスへ向けて～

国立大学法人東海国立大学機構 名古屋大学糖鎖生命コア研究所統合生命医科学糖鎖研究センター・名古屋大学大学院医学系研究科総合保健学専攻の松井 佑介 准教授の研究グループは、愛知がんセンターの阿部 雄一 主任研究員、東京医科歯科大学の宮野 悟 教授との共同研究で、がん発症の原因と考えれるタンパク質複合体を、超高速かつ網羅的に解析することが可能な、データ解析手法を新たに開発することに成功しました。

本研究により、これまで研究されてきたがんゲノムの異常が、その実体であるタンパク質レベルにおいて、がんの発生や転移、薬剤耐性とどのように関係しているのかを網羅的に明らかにし、新たな薬剤標的の発見へと繋げていくことが期待できます。

本研究成果は、2021 年 9 月 17 日付国際雑誌「Bioinformatics」に掲載されました。

【ポイント】

・タンパク質複合体[注1)]の制御異常を予測するデータ解析手法を開発した。
・本手法は、質量分析特有の測定時に生じる欠損値やノイズに対して頑健である。
・計算の高速化も実現し、計算時間は従来と比べて最大 5000 倍向上した。
・110 人の腎癌検体プロテオームデータ[注2)]へ応用し、既知のゲノム変異や薬剤的標的を含む新たなタンパク質複合体異常を 200 以上同定した。

【研究背景と内容】

がんは、ゲノム変異だけでなく、エピジェネティック[注3)]および転写制御を含む、様々な分子異常によって引き起こされる複雑な疾患であることが知られています。しかし、それらのイベントが、プロテオームのレベルにおいて、どのように特徴付けられているのかは、ほとんど解明されていません。個々のタンパク質は、タンパク質複合体を形成して機能することが知られています。がんなどの疾患患者の細胞では、ゲノム異常などにより、複合体が正常に機能しなくなると考えられますが、そのメカニズムは十分にわかっていません。そのため、タンパク質複合体ではどのような異常が起こっているのかを解明することは、治療戦略の策定や薬剤標的の予測において重要です。

そこで、本研究では、液体クロマトグラフィー質量分析器(LC/MS/MS)[注4)]を用いて測定されるタンパク質発現データに基づいて、共発現変動解析[注5)]という枠組みを用いることで、タンパク質複合体においてがん特異的な異常を予測するための、データ解析手法「Robust Differential Co-Expression analysis (RoDiCE)」を開発しました（図 1）。「RoDiCE」の特徴は、測定に由来するノイズや修飾語翻訳等による外れ値などノイズに頑健であることです。また従来の方法では、あるタンパク質と他方のタンパク質の共発現構造を群間で比べるというペアワイズな方法が主でしたが、本手法では、複数のタンパク質間の共発現構造を同時に比較することが可能であり、また線形的な共発現構造のみならず、非線形な共発現構造を捉えられるため複雑な構造変化も検出することができます。実際の腎癌患者 110 名のデータへ適用したところ、主要な薬剤標的遺伝子を含む 200 以上の複合体が検出することができました（図 2）。今後、様々ながん種における複合体異常のメカニズムの解明や薬剤標的のスクリーニングにも役立つと考えられます。

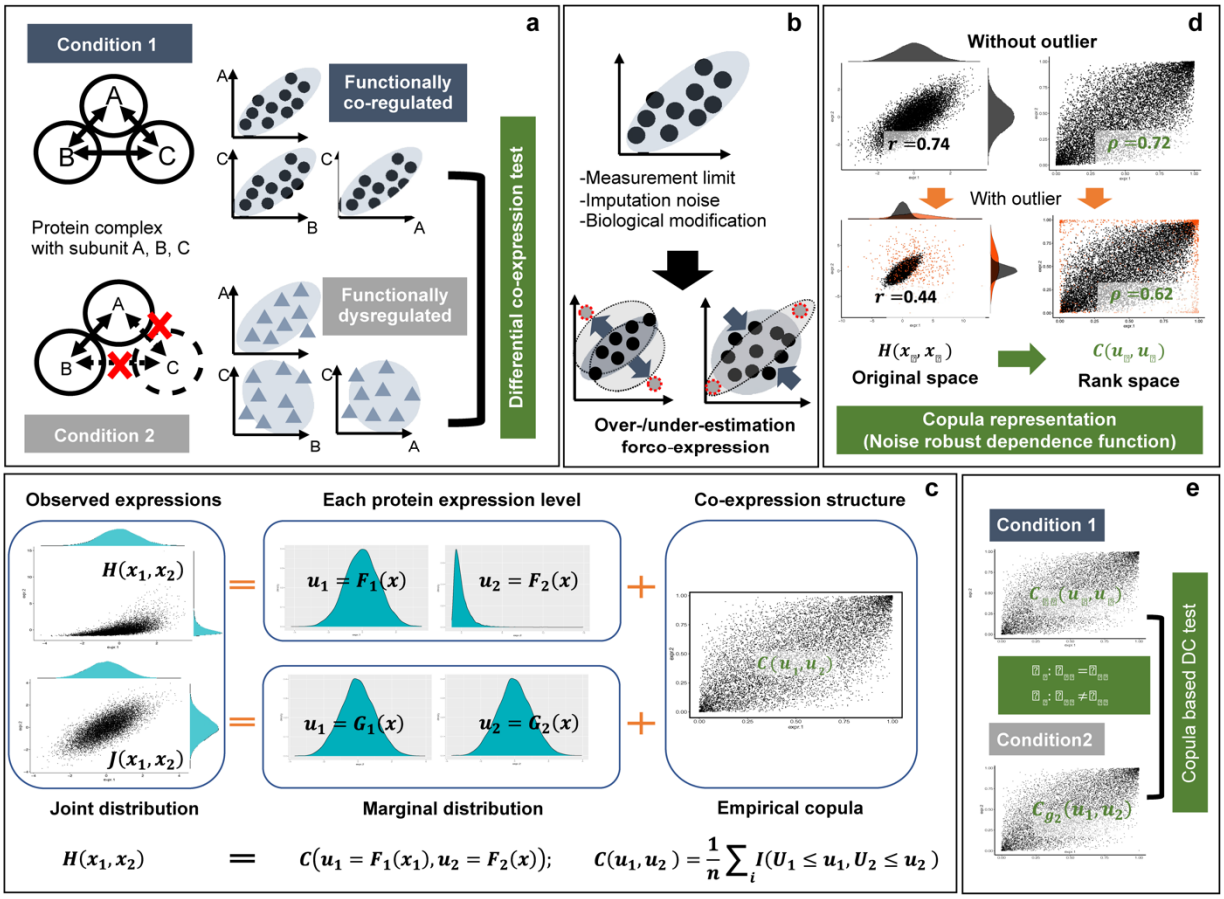

図 1. RoDiCE の概要

a)RoDiCE による解析目的：2 つの異常なグループを比較することで、異常なタンパク質複合体を特定することが目的である。 **b)タンパク質共発現と異常値**：LC/MS/MS で測定されたタンパク質の発現量には、複数の原因によるノイズが加わることで、共発現構造の過大評価（または過小評価）の原因となっている。 **c)コピュラ**[注 6)] **によるタンパク質発現量のモデル化**：RoDiCE モデルでは、観測されたタンパク質発現の多変量分布を、各タンパク質の振る舞いを表す周辺分布と、タンパク質間の潜在的な共発現構造を表す経験的コピュラ関数に分解する。 **d)コピュラの頑健性**：コピュラは、データのスケールをランク変換することで外れ値に対する頑健性を実現。 **e)共発現変動解析**：RoDiCE ではコピュラの比較に対して並び替え検定を行うことで共発現構造の変化を推定している。

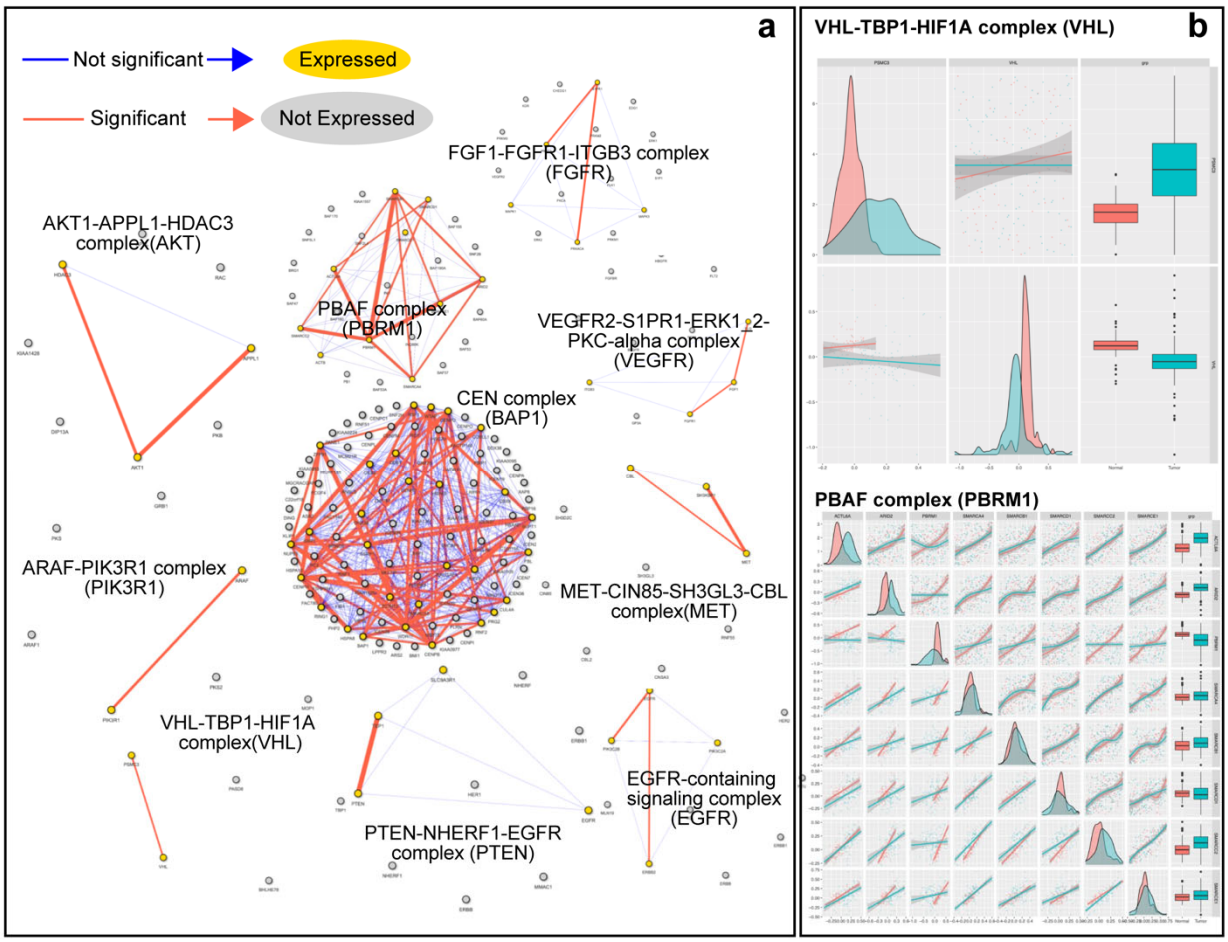

図 2. ドライバー遺伝子に関連するタンパク質複合体を同定

a) 既知のドライバー遺伝子および薬剤標的遺伝子における共発現変動したタンパク質複合体。赤は、タンパク質複合体のサブユニット間の共発現に差があるペアを示す（5%有意水準）。線の太さは-log10(p-value)。青い線は有意でないペアである。黄色のノードは本研究で実際に LC/MS/MS で発現を測定したタンパク質、灰色のノードは測定されなかったタンパク質を表す。 **b)共発現構造を持つ VHL-TBP1-HIF1A 複合体と PBAF 複合体の例**。青は腫瘍群、赤は正常群を表し、対角線上にタンパク質発現の密度分布を示した。下段の対角線上には、コピュラ変換前の共発現パターンが示されており、上側の対角線上にはコピュラ変換後の共発現パターンが示されている。

【成果の意義】

これまでにも共発現変動解析は、マイクロアレイ[注 7)]や RNA-seq[注 8)]を用いた研究において、異なる条件間の共発現構造の違いを検出するための標準的な手法として用いられていました。しかし、LC/MS/MS に対しては、翻訳後修飾による外れ値の存在や、計測限界によって生じる大量の欠損値を処理することによるタンパク質発現量の誤差などによって、これまでの遺伝子発現解析よりも大きなノイズを有することが知られています（図 3）。

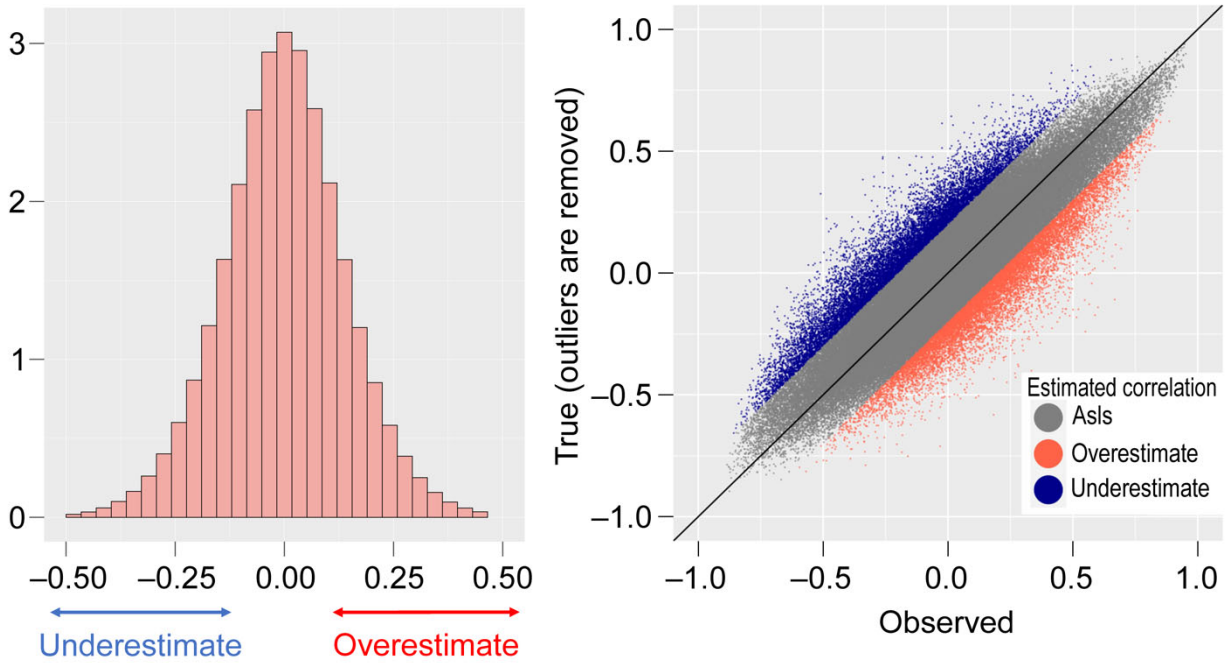

図 3. 共発現に対する外れ値の影響の例：外れ値のサンプルを除去する前と後のピアソンの相関[注 9)]の差異。左のパネルは、相関の差異のヒストグラム。右側のパネルは、外れ値のサンプルを除いたものに対する元の相関の散布図。

またタンパク質複合体では、複数のタンパク質間の共発現構造が同時多発的に破綻することから、複合体ごとの共発現構造の同時変化を検出する必要がありました。

本手法ではこれらの課題に対して、コピュラ(Copula)と呼ばれる確率モデルを応用したデータ解析手法を開発することで解決を目指しました（図 1)。タンパク質複合体を構成する複数タンパク質の発現量を、コピュラを用いてタンパク質間における共発現構造（コピュラ関数）と、各タンパク質の発現量（周辺分布関数）を表す構造に分解し、ノイズの影響が顕著に現れる各タンパク質発現のふるまいを共発現構造から切り離すことで、ノイズや欠損値に対して頑健な共発現構造の比較が可能となりました（図 4)。また、コピュラは複数の（線形・非線形な）依存関係を一度にモデル化することが可能であるため、タンパク質複合体を構成する複数タンパク質の共発現構造の同時多発的な構造破綻を安定的に検出できます。

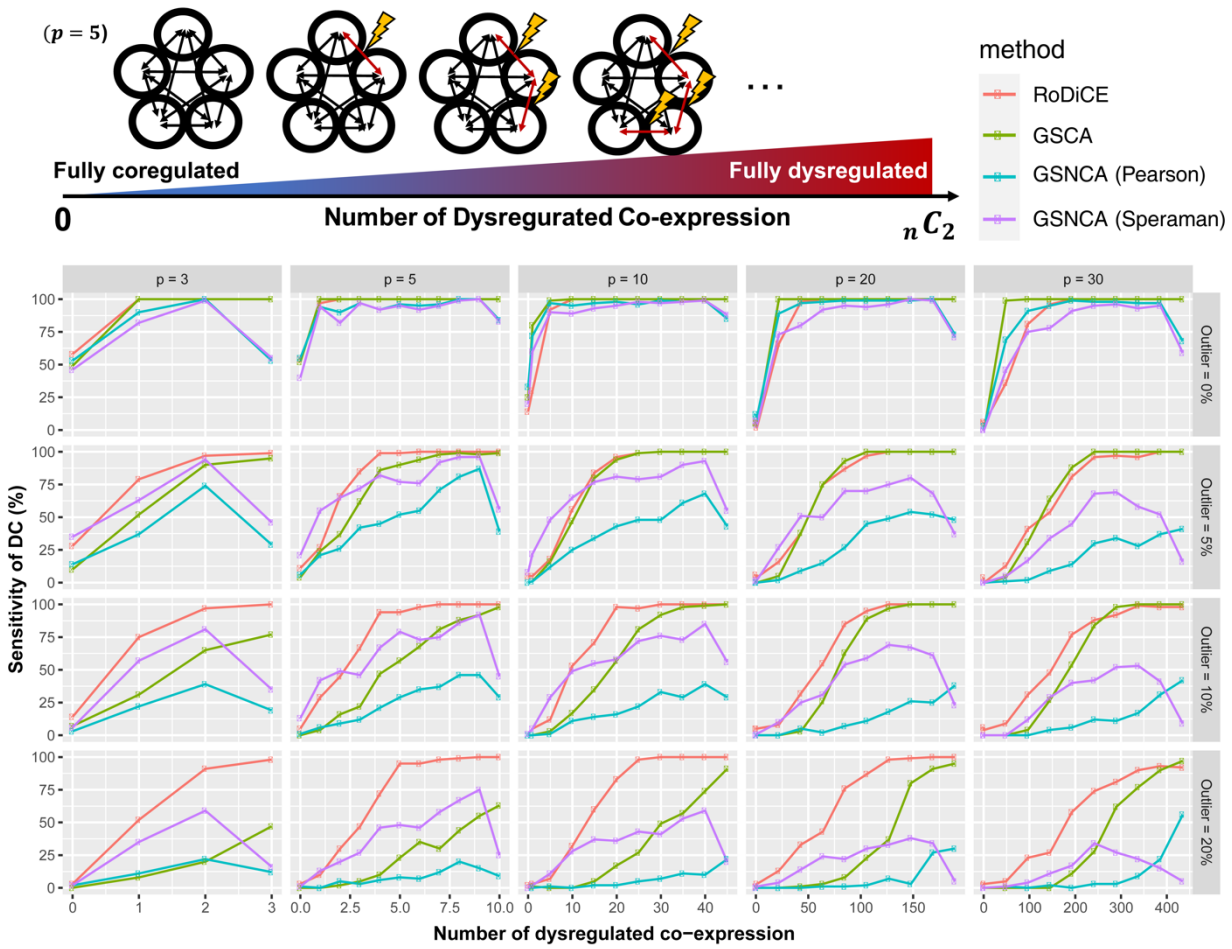

図 4. 感度と外れ値の割合：横軸に外れ値の割合をとり、縦軸に各手法による共発現の違いの感度（有意水準 5%）を示す。

しかし、実際の大規模プロテオームデータへの応用にあたっては計算時間も大きな課題でした。そこで本研究では実際の大規模プロテオームデータへと応用するために、高速化するための工夫をすることにより、高速化前と比較して最大で 5000 倍の高速化に成功しました。これにより大規模集団であっても現実的な時間内でタンパク質複合体異常を網羅的かつ安定的に予測することが可能です（表 1）。

表 1： 計算速度（10 回の平均、単位は秒）

	p = 2	p = 3	p = 5	p = 10	p = 20	p = 30
高速化後	**0.373**	**0.457**	**0.499**	**0.605**	**0.771**	**0.986**
高速化前	25.301	51.482	125.753	437.747	1674.990	5101.482

【用語説明】

注 1）タンパク質複合体:

複数のタンパク質が集まったグループのこと。個々のタンパク質はタンパク質複合体を形成することで、細胞内における多くの生物学的プロセスに関与している。

注 2）プロテオームデータ：

細胞内で働いているタンパク質全てに関する情報のこと。Protein（タンパク質）+ ome（全ての）に由来する造語

注 3）エピジェネティック：

細胞では DNA 配列を変えずにその形態や機能を調節する仕組みがある。それらのメカニズムを総称してエピジェネテクスと呼んでいる。

注 4）液体クロマトグラフィー質量分析器(LC/MS/MS)：

細胞内で働いているタンパク質全てを一度に検出して定量化する測定技術。

注 5）共発現変動解析：

タンパク質同士が共に働いていると考えられる状態を共発現と呼ぶ。タンパク質 A とタンパク質 B があったとして、細胞内において A が増えると同時に B も増えのであれば、A と B は互いに協調関係にあると考えられる。共発現変動とは、疾患群と健常群で比較した時に、健常群では共発現していたのに、疾患群ではそれが失われて共発現が変化していることを共発現変動と呼ぶ。

注 6）コピュラ：

個々の確率的な振る舞いにおける互いの依存関係を表すための確率モデル。コピュラを用いることで極めて多様な共発現を捉えることができる。

注 7）マイクロアレイ：

一度に細胞内にある多数の遺伝子の存在量を測定する技術である。予め設計・固定した DNA/RNA が基板上に整然と並べられており、試料中に類似した配列を有した遺伝子があると、その存在量に応じた蛍光強度として測定値が得られる。

注 8）RNA-seq：

一度に細胞内にある全ての遺伝子の存在量を測定する技術。人のゲノムを構成する 30 億の ATGC からなる文字列を一度で大量に読み取ることができるシークエンサーと呼ばれる機械を用いて、遺伝子の存在量を表す配列断片を一度に大量に読み取る。

注 9）ピアソンの相関：

異なる二つの対象間の依存関係を捉える指標。例えば、タンパク質 A とタンパク質 B において、タンパク質 A の存在量が増えると、タンパク質 B が比例して増えていく

関係にあるほどピアソンの相関は大きくなり、逆に関係がなければゼロに近づく。

【論文情報】

雑誌名：Bioinformatics
論文タイトル：RoDiCE: Robust differential protein co-expression analysis for cancer complexome
著者：Yusuke MATSUI, Yuichi ABE, Kohei UNO, and Satoru MIYANO
DOI:10.1093/bioinformatics/btab612
URL:https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btab612/6371291