総説

多層的疾患オミックス研究

大保木啓介† 松本健治 斎藤博久

IRYO Vol. 64 No. 3 (135－145) 2011

要旨 生体構成物質についての近年の網羅的解析手法のめざましい発展により，さまざまな生体物質についての網羅的データを取得し解析すること（いわゆる多層的オミックス/バイオモレキュロミクス解析）が可能となりつつある．本稿では，各種生体物質の網羅的解析技術（高密度マイクロアレイ、高速シークエンサー、質量分析計）とその応用面に加え，網羅的データの問題点について触れ，将来期待される各種の物質群の情報からなる多層的なオミックスデータの可能性について述べる．

キーワード 多層的オミックス解析，高密度マイクロアレイ，高速（大規模並列）シークエンサー，質量分析計，バイオモレキュロミクス

はじめに

細胞レベルや分子レベルまで拡大していくと，生体はさまざまな有機化合物，つまりDNAをはじめ，RNA，タンパク質，脂質，糖鎖などの生体有機化合物によって構成されていることがわかる．近年，とくにヒトゲノム配列が最初に解読[1]されて以降，これらの生体物質を網羅的に解析するための技術（高密度マイクロアレイ，大規模並列シークエンサー，質量分析計）が発達してきた．装置や解析費用が高価であるために一般化のスピードは遅いが，そう遠くない将来，ありとあらゆる生体分子についての網羅的解析から導かれた情報（Biomoleculome：バイオモレキュローム）を利用することが生物学研究，あるいは臨床研究の基本ツールになると考えて差し支えない．人間やモデル動物，有用植物，微生物などを構成するバイオモレキュローム研究を推進することによって，きわめて複雑な生体分子の振る舞いを理解することが可能になると考えられ，病気の原因解明，新規な医薬品の開発などに大きな進展が期待されている．現在NatureやScienceに掲載される研究には，網羅的データを使った発見型研究や，新しく開発した網羅的データ取得方法を用いた研究が大変多い．これは，網羅的データが今まで観察できなかった新しい型のデータであって，生物学を全く新しい観点から捉える力を持つこと，さらに

国立成育医療研究センター研究所 免疫アレルギー研究部 †その他（研究員）
別刷請求先：大保木啓介 国立成育医療研究センター研究所 免疫アレルギー研究部
〒157-8535 東京都世田谷区大蔵2-10-1
（平成22年9月21日受付，平成23年3月11日受理）
Exploratory Study of Diseases by Multi-Omics Data
Keisuke Oboki, Kenji Matsumoto and Hirohisa Saito, National Research Institute for Child Health and Development, Department of Allergy and Immunology
Key Words: multi-omics analysis, high density microarray, high-speed (massively parallel) sequencer, mass spectrometer, biomoleculomics

は医学判断を根本的に変える可能性を見据えての掲載傾向であろう．網羅的かつ大量のデータを定量的に扱うシステム生物学と呼ばれる研究分野も確立されつつある．しかし，生物学に携わる研究者あるいは臨床医が本来の守備範囲に加えて情報処理技術をマスターし，システム生物学者になるのは多くの場合現実的ではない．今後はむしろ，網羅的生体情報を適切に扱うことを素養の一つとすべき傾向が強くなっていくだろう．ただし，大量データから意味ある情報を抽出しようとする場合，大量の情報を正確かつ効率的に扱う情報処理技術と，生物学的解釈を行うための生化学，分子生物学，細胞生物学，遺伝学，統計学，臨床医学にいずれも通じた医学・生物学者による「解釈」が必須となる．これらすべてを個人または少数の研究グループで処理しなければならないという新しい時代に突入しつつある．扱うデータの質と量の変容と共に，医学・生物学研究の体制も変容していく可能性がある．

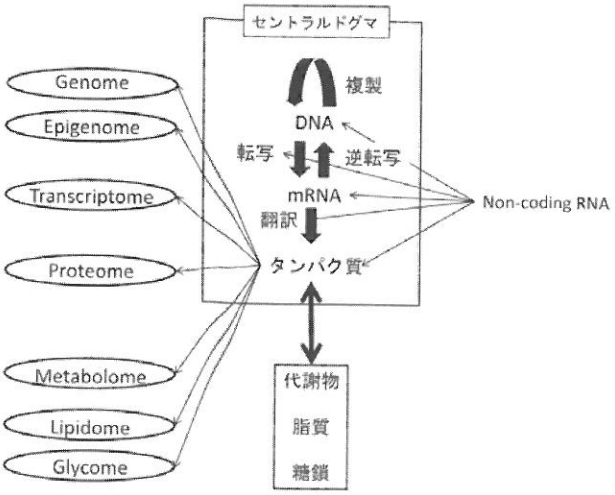

図1　セントラルドグマとオミックス解析
DNA から RNA を介してタンパク質が生合成されて細胞機能を成立させるとするフランシス・クリック提唱のセントラルドグマに，現在までの知見を加えた生体機能分子群の関連図．それぞれの物質群ごとの包括的データ（～ome）の解析をオミックス解析と呼ぶ．エピゲノミクスは DNA のうちシトシンのメチル化，およびヒストンタンパク質（DNA を巻き付け染色体構成に預かる）の N 末端の特定アミノ酸残基の化学修飾変化（メチル化，アセチル化など）を測定する．近年その存在が明らかになってきた non-coding RNA は低分子 RNA（miRNA 等）を含み，その機能についての研究が進行中である．

要素だらけの医学，生物学

私たちの体を構成する細胞は核を有する細胞であり，細菌などの原核生物の細胞形態とは大きく異なる．真核生物は哺乳類を含む多種多様な生物集団を形成していて，その最もシンプルなモデルとして強力な遺伝学とともに研究が進んでいるのが酵母である．酵母をツールとした近年のノーベル賞には細胞周期遺伝子の発見（2001年）がある．今や細胞周期研究は癌研究の基盤の一つとなった．2009年のカナダのガードナー賞（ノーベル賞受賞者が多い）を，iPS 細胞[2]で山中伸弥博士が受賞している．あまり報道されていないが，このとき同時に，京都大学・森和俊博士と UCSF・Dr. Peter Walter による酵母研究から見いだされた小胞体ストレス応答の研究[3)4)]が受賞している．小胞体ストレス応答は，酵母の研究から哺乳類の研究へと発展し，糖尿病や神経変性疾患をはじめ，多様な疾患に関わることが明らかとなっている．酵母はゲノムサイズが小さく（約1,200万塩基），モデル真核生物のなかでは最初期にゲノム配列も解読されており，タンパクをコードする遺伝子はほぼ完全に明らかになっていて，読み取り枠（open reading frame：ORF→アミノ酸コード領域ではないかと予想される DNA 領域のこと）数は6,607個である．しかし，遺伝子（ORF）の存在が明らかになることと，その機能が明らかになることとの間には，大きなギャップがある．酵母で同定された遺伝子のうち，機能未知の ORF が1,724遺伝子存在しており，実に全体の25％を占める．モデル生物の酵母では，早くから体系的にすべての遺伝子の破壊株が作成されてきたが，いまなおその機能が明らかでない遺伝子が数多く存在する．

それではヒトではどうだろうか．まず，ヒトの生物学的モデルとしてよく使われるマウスについてみてみたい．サルがヒトに最も近縁であるのはいうまでもないが，マウスは犬猫や牛馬よりも進化上ヒトに近く[5)]，モデル生物としてマウスはわれわれのパートナーといってよい．マウスには25,878個の protein-coding gene があり，このうち11,259個（約４割）の遺伝子機能が「わかっている」とされる（Mouse Genome Informatics：http：//www.informatics.jax.org/）．すなわち，（驚くべきことに）およそ６割は機能不明であると考えて差し支えない．ヒトの最新版ゲノム情報（GRCh37，Feb2009）では protein-coding gene が21727個（Ensembl：http：//uswest.ensembl.org/index.html）である．マウス

とヒトの遺伝子は非常によく似ていて，両者は遺伝子機能についての知見が補完関係にあるので，実質的にヒトでも６割程度が機能不明であると考えてよい．最近明らかになりつつある，多種類の「謎の」非コード型RNAは上記２万数千個には含まれていない．セントラルドグマ（図１）的見地からタンパク質をコードする遺伝子に限ってみても，全体の６割の遺伝子機能がよくわからないまま，網羅的なデータがとられているというのが実情なのである．つまり網羅的データとは，機能のわからない要素を多く含む，言い換えれば生物学的解釈ができない要素を多く含む解析手法であるといえる．タンパク質の特徴的な配列情報（モチーフ）から機能がある程度推測される場合もあるし，よく練られたRNA網羅的発現データ解析から特定の遺伝子機能を予測することも可能であるが，それらはあくまで推測に過ぎず，実験的に機能が確認あるいは「発見」されるまでは明らかとなったことにはならないことに注意したい．NCBIなどのデータベースをみると，「同定されている」遺伝子が多いが，それらの多くの機能はいまだに未解明である．

知らないものはみえない

診断であれ，研究であれ，さらにいえば日常生活であっても，われわれがみて認識することができるのは注意の働きのおかげである．この注意の働きは「対象を知っていること」によって大きく促進され認識の助けとなる．だまし絵の例を考えるとわかりやすい．だまし絵の中にある隠された対象物は，初見ではなかなか見いだしにくい．しかし，いったんそこにその「対象物がある」ことを認識し，記憶すると同時に，そのだまし絵はだまし絵ではなくなり，その対象物は容易に注意を引き識別されるようになる．病名と病態を知らなければ診断は不可能なように，分子の機能とその生物学的背景を知らなければ，網羅的データから生物学的意味を抽出することは不可能である．すべての要素の機能が明らかになるまで，今後も生化学的，分子生物学的に単一分子の機能を解析し考究することの重要性は変わることがないというのが論理的帰結である．繰り返しになるが，網羅的データは要素の性質についての知識が基盤である．網羅的データは2010年時点でも機能のわからない要素を多く含んでいる．

ヒトゲノムプロジェクトの終了とその後

米国NIHの現所長であるFrancis Collinsは国際ヒトゲノムプロジェクトをJames Watsonのあとに率いた人物である．現在の医学・生物学を一変させるゲノム配列解明という取り組みを実質的に完遂した人物であるという評価は疑いようがない．ヒトゲノムプロジェクトがヒトゲノム配列解析データを世に出した後，続いてCollinsがどのようなプロジェクトを推進しているのかが大変興味深い．彼がヒトゲノム解読後取り組んだのは，1000人ゲノムプロジェクト（1000人のゲノム配列を解読する），だけであれば話はわかりやすい．しかし，ほかにも重要なプロジェクトを立ち上げている．その一つが米・加・欧州からなる国際的マウス遺伝子ノックアウトプロジェクトであり（http://www.knockoutmouse.org/）[6]，米国では2006年から予算措置がとられている．先に述べたように，マウスで約６割のORFがコードするタンパク質の機能は不明である．たとえばある細胞のトランスクリプトーム情報を得たとしても，約６割の要素からは１次情報が何も引き出せず，網羅的情報の多くは塩漬け状態となるのが現状である．Collinsは遺伝子の機能を明らかにすることがポストゲノム時代に必須であり，その有力なモデル動物にはヒトに最も近縁なマウスを使い，多数の機能未知の遺伝子をノックアウトするプロジェクトが必要であると考えたのである．機能の明らかとなっていない遺伝子の働きを明らかにするために，そのnull変異体（遺伝子機能を欠失させたもの）を得るのは，すべての生物解析の常套手段であり，最もパワフルな手法である．残念ながら，日本はこのプロジェクトに公的には全く貢献できていない．遺伝子ノックアウトは専門性の高い技術であり，１遺伝子当たりのノックアウトマウス作成作業には非常に時間がかかる．該当する遺伝子が発生に必須であったり，生存に必須の遺伝子である場合には，null変異体は致死性となってしまうので，コンディショナルノックアウトマウス（任意の臓器や，任意の時期に遺伝子を欠失させる）を作成しなければならない．続く，ノックアウトマウスを使用した疾患モデル解析がプロジェクトの核心であることも明らかで，長期間にわたる地道な取り組みが必要である．日本においても，単純かつ重要であるこのような研究が強力に推進されることが，ヒトにおける網羅的データを使ったオミックス研究成果を有効たらしめる．

オミックス解析のためのハードウエア

生体分子の網羅的解析を可能としている基盤技術について概説する（図２）.

１．高密度マイクロアレイ

メッセンジャーRNA(messenger RNA：mRNA)，マイクロRNA(micro RNA：miRNA)（後述），non-coding RNA（後述）の発現量や，DNAのコピー数多型，DNAメチル化解析などに用いられる．高密度マイクロアレイは，ガラスやシリコンの小さな基板上に，すでに同定されている遺伝子の相補鎖オリゴヌクレオチドや，ゲノムDNAを高密度にスポットしたものになる．従来のDNA解析としてのサザン法やRNA解析のノーザン法では，フィルターに固定した全DNAあるいはRNAに対して，放射性同位元素でラベルした検出したい遺伝子プローブをハイブリダイゼーションして目的の（おおよそ１種類の）遺伝子を検出していた．高密度マイクロアレイでは，逆に，基板上に前もって既にわかっている遺伝子情報を元にしたプローブを数万種類単位で固定してしまい，そこに蛍光ラベルした細胞由来の核酸をハイブリダイゼーションする．スポットされている個々のプローブの位置がわかっているので，あとはスポットごとの蛍光強度を高感度スキャナで測定すれば，DNAあるいはRNAの量に相関する大量のシグナルを得られる仕組みである．プローブオリゴヌクレオチドの基板上合成技術やスポット技術の向上によって，精度が高く，高密度に，より長いプローブをスポットすることが可能となりつつあり，シグナルダイナミックレンジが従来のものより100倍ほど向上し，特異性も向上しつつある．データ解析はメモリを増設した通常のパソコンであればストレスなく解析が可能で，安価な国産ソフトウエアも出始めている．

２．高速（大規模並列）シークエンサー/次世代シークエンサー

核酸の塩基配列を決定する装置をシークエンサーと呼ぶ．従来の汎用シークエンサーはサンガー法（ダイデオキシ法）と呼ばれる方法を採用しており，現在も世界の多くの研究室で使用されている．ヒトゲノムプロジェクトをはじめとする多くのゲノムプロジェクトはこの方法で解読されてきた．サンガー法では，読み取りたい配列を持つ核酸を鋳型として一方向のDNAポリメラーゼ反応を行い相補鎖を合成する．このとき反応液中には通常の４種の核酸に加えて，蛍光ラベルした４種の核酸アナログを混合して反応を行うことで，オリゴヌクレオチドの3'端に核酸ごとに区別できる蛍光を持った数十～数百塩基の断片を１塩基違いで作成できる．これをゲル電気泳動で短い断片順に分離して，順々に蛍光を読み取っていく方法がサンガー法である．サンガー法で読み取れる長さは500-600塩基程度である．これに対し，高速（大規模並列）シークエンサーでは，

図２　オミックス解析のハードウエア
包括的な生体分子情報を測定するための代表的な機器．▽マイクロアレイはRNAの発現量測定，あるいは，コピー数多型，一塩基多型（SNPs），DNAメチル化測定ツールとして一般的に普及している．▽高速（大規模並列）核酸シークエンサー．これまでのサンガー法とは異なり，大規模な並列読み取り法を採用したことから次世代シークエンサーとも呼ばれる．高速シークエンサーによって大量の核酸配列の読み取りが短時間で可能となった．前処理を施すことで他の核酸関連解析（ChIP-seq，RNA-seq等）も可能である．米国メーカーが開発競争を繰り広げており，PCRを用いない新しい核酸読み取り原理による超高速シークエンサーも販売開始予定である．▽質量分析計は低分子化合物測定機器としての歴史が長い．微量試料の分離技術，質量測定技術等の革新・開発によって大量の質量同定が可能となった．さらに高分子の測定が可能となったことから，タンパク質や糖鎖についても大量（網羅的）同定が行われている．

ごく短い断片（現在は50–150塩基程度）を大量（数千万）かつ同時に配列決定する．具体的には，ランダムに基板上に貼り付けた解析対象の断片化DNAを鋳型として，蛍光標識核酸を用いた合成を「基板上」で行ったあと，大量のスポットの蛍光をスキャンした画像を記録する．この合成→記録というプロセスを繰り返して，各スポットごとに蛍光変化を追うことで，配列を短時間に大量決定する（sequence by synthesisと呼ぶ）．これは，すでに明らかとなっているゲノム配列があることを利用し，新しく読み取った短い配列断片でも既知のゲノム配列に当てはめていくことが可能であることを利用しているといえる．新たなゲノムの配列決定も可能であるが，応用的用途に向いており，実際さまざまな用途がある（mRNA-seq，miRNA-seq，ChIP-seqなど，後述）．配列読み取り方法として新しい技術が米国を中心に試されており，高速シークエンサー装置の技術革新が現在進行中である．読み取り塩基数の延長や読み取り方法，読み取りスピードがさらに向上し，コストダウンできる超高速シークエンサーが登場すれば，ひとりの人間のゲノムを安価かつ短時間で読める時代に突入し，核酸の解析手法は大きく変化すると考えられている．一方で，少なくとも現行高速シークエンスは，画像データを利用すること，非常に短い断片配列が大量に出るという性質から，一度の読み取りデータ量は数テラバイト（塩基配列としては数十億塩基）に及ぶ．ゆえに高速（大規模並列）シークエンサーによる解析には，コンピュータの高処理能力，大規模な記憶容量，情報処理技術者が欠かせない要素である．一塩基当たりのランニングコストはサンガー法よりはるかに安価であるが，1ラン当たりの解析単価は現時点で非常に高価である．

3．質量分析計

質量分析をするための機器を質量分析計（mass spectrometry：MS）と呼ぶ．どのような生体物質でも測定可能で，生物学分野ではタンパク質，脂質，低分子代謝物などに適用されることが多い．MSはどれも物質をイオン化し，移動させるとき質量が大きいと移動速度や移動距離が短くなる原理を使って測定する．誤解を承知で簡約すれば，クロマトグラフィーや電気泳動などと同じ原理である．分析する試料をイオン化させ，電磁気力により質量ごとの差をつくり，イオンの質量を分析する．イオンは，純粋な質量としてではなく，質量数（イオンの質量）/電荷数（イオンの価数）＝m/z（質量電荷比）に従って分離され，データ出力される．MSは大まかには，試料導入部，イオン源，分析部，イオン検出部から構成され，試料はこの順に運ばれる．生成したイオンを安定に検出器まで到達させるため，イオン源，分析部，イオン検出部は真空状態に保たれる．出力された質量ピークはPCによるデータベースとの照合により物質同定される．タンパク質および生体代謝分子については，個々の分子数の多いものと少ないものを比較すると，一説では10^7ほどの開きがあるという．現行機種の測定ダイナミックレンジは10^3から10^4であり，10^7は現在技術的に実現困難であるが，サンプルの分画法やカラム分離技術と質量分析計の技術改良により，かなりの網羅的な同定が可能となった．

試料導入部は多様な物質の混合物である試料を装置内に導入するいわゆる前処理部位である．試料の性質により導入法が異なり，高速液体クロマトグラフィー（HPLC/UPLC）や，キャピラリー電気泳動（CE）等を質量分析計に直結し，微量の移動相を連続的に導入する技術が網羅的データには欠かせない（LC/MS（エルシーエムエスまたはエルシーマス），CE/MS（シーイーエムエスまたはシーイーマス）など）．実際の質量分析現場で最も幅広く用いられている方法がLC/MSである．

イオン源は試料物質に電荷を持たせる部位である．イオン化のさまざまな手法が開発されている．マトリックス支援レーザー脱離イオン化（Matrix Assisted Laser Desorption Ionization：MALDI）法は，試料を芳香族有機化合物などのマトリックスと混ぜて結晶を作り，これにレーザーを照射することでイオン化する方法である．MALDIイオン化法によるタンパク質のイオン化に初めて成功した島津製作所・田中耕一が2002年ノーベル化学賞を受賞した．エレクトロスプレーイオン化（Electro Spray Ionization：ESI）法は，主にLC/MSにて使用される方法で，キャピラリに高電圧を印加すると試料溶液が自ら噴霧，イオン化する現象を利用した方法である．MALDIと同じく，高分子化合物のイオン化に特に優れる．高速原子衝突（Fast Atom Bombardment：FAB）法は，試料をマトリックス（グリセリンなど）に混ぜ，ここに高速で中性原子（Ar, Xeなど）を衝突させることでイオン化する方法である．試料を気化する必要がないため，広範囲の物質に使用できる．

分析部，測定部は，イオン化された試料を分離，同定する部位であり，質量電荷比（m/z）の近接ピークを区別する分解能，マスレンジと呼ばれる質量測定可能範囲，そして，検出部の性能に依存するシグナルダイナミックレンジがここで決まる．分離の方法は多種類あり，磁場偏向型，四重極型，イオントラップ型，飛行時間型，フーリエ変換イオンサイクロトロン共鳴型などがあり，それぞれ特性，装置価格ともさまざまである．タンデム型と呼ばれる質量分析計があり，上記の分析部を複数・直列に組み合わせたもので，MS/MS（マスマスあるいはエムエスエムエス）と呼ぶ．MS/MSでは，一度質量測定したイオン化物をさらに断片化して質量測定することができるため分解能が向上する．

網羅的解析対象となる生体物質

各生体物質ごとに行われるオミックス研究と，疾患研究への貢献のポテンシャルを概説する．よく知られているように，生体物質に-omeを付けるとその包括的情報を指し，-omicsを付けると包括的情報解析研究を指す．

1．核酸

1）ゲノミクス（genomics）

核酸のうちDNA配列の包括的情報研究をゲノミクスと呼ぶ．すべての基盤となるヒトゲノム配列は国際プロジェクトにより公開されており，継続的なメンテナンス，付随情報の追加等が行われている．

SNP（small nucleotide polymorphism）：疾患原因の探索としてDNAを解析する場合，単一遺伝子変異からなる遺伝病であれば，古典的にDNA多型マーカーを使用して家系解析を行えばよいが，生活習慣病，アレルギー疾患，自己免疫疾患等は原因となる遺伝子が多因子であることが多く病因解析は長年困難であった．一つのヒトゲノムセットが30億塩基対であることから，効率のよいゲノムDNA情報の使用法としてさまざまな工夫が行われており，近年進展が著しい方法の筆頭がゲノムワイド関連解析（GWAS）である．ゲノムDNAのほとんどは種内で同一であるが，個体間で多型を見いだせる部分も存在し，その代表的な多型がSNPと呼ばれる．ヒトのSNPをデータベース化する国際プロジェクトが進行し，この多型を高密度マイクロアレイ等によって網羅的に検出することができるようになった．患者と健常者のDNAを数千人規模で集めることができれば，このSNPを使うことで，GWASを行うことができ，原因のわからなかった疾患の患者に特異的に連鎖するSNPを統計的に抽出し，疾患原因遺伝子の同定が可能である．

コピー数多型（copy number variation：CNV）：ゲノムDNAの多型の一つに，コピー数多型CNVがある．細胞1個当たりの染色体1セットのうち，常染色体であれば2コピーあるはずであるが，重複や欠失などで3コピーであったり，1コピーである場合がある．このゲノム領域をCNVと呼ぶ．CNVは直接の疾患原因にもなるが，実際には健康なヒトのゲノム中にも高頻度にみられる多型である．CNVは高密度マイクロアレイによって効率よく検出可能である．SNPよりも大きな領域を含むため，多型マーカーとして多様な用途と，疾患原因解明への直接的な貢献も期待される．

2）トランスクリプトミクス（transcriptomics）

生体内のRNA配列の包括的情報研究をトランスクリプトミクスと呼ぶ．

マイクロアレイによるmRNA解析：タンパク質をコードするRNAについて，どのような転写産物があるのかが長年研究され，転写産物のカタログ化・データベース化が現在も進められている．この情報を利用して，高密度マイクロアレイ法により包括的転写産物の相対定量を行う方法が普及している．これを用いれば，疾患組織と健常組織の転写プロファイルを比較することができたり，刺激前後の包括的時系列データなどを取ることができる．癌などの疾患における治療反応性や予後予測としてのプロファイルデータとしても有用である．

高速（大規模並列）シークエンサーによるmRNA解析：ヒトのタンパク質コード遺伝子（protein-coding gene）は21,000個ほどであるが，そこから転写されるRNAは，プロモーターが異なったり，スプライシングがあったりするために，実際には非常に多様な分子集団を形成する．高速（大規模並列）シークエンサーであれば，それらを区別し，実際にどのようなアイソフォームが存在するのかが明らかになる．また，大量に読み取りを行えば，読み取り数をカウントすることで，デジタルな遺伝子発現量データを得ることができる．加えて，白血病細胞を中心に，別々の遺伝子に由来するmRNAが，転座などで融合遺伝子あるいはキメラmRNAを形

成していることが報告されている[7]．古くは慢性骨髄性白血病（chronic myelogenous leukemia：CML）のフィラデルフィア染色体でのBCR-abl融合遺伝子[8]が有名であるが，癌および他の疾患でもまだ未同定かつpathogenicな融合遺伝子が存在していると予想され，高速（大規模並列）シークエンサーによる発見的配列解析（RNA-seq）が期待されている．

non-coding RNAとmiRNA：タンパク質をコードしないタイプのRNAが細胞内には多種類存在する．古典的にはリボソームRNA，トランスファーRNA，スプライシングなどを制御するsnRNPを構成するsnRNA，あるいは核小体に局在するsnoRNAなどである．近年になり20-25塩基の翻訳調節性のマイクロRNA（miRNA）が存在することや，poly Aを持つmRNA様non-coding RNAが予想よりも大量に転写されていることが明らかとなった．mRNA様non-coding RNAの一部は大型介在性非コードRNA（lincRNA）と呼ばれはじめている．miRNAはすべてが同定されているわけではないが，1000種類以上あると考えられている．理研・林崎グループによる解析から，non-coding RNAはmRNAよりも多種類であると考えられているが，遺伝子量補正に働くXistなどのごく一部を除いて，ほとんどの機能は明らかではない[9]．疾患マーカーとなる可能性を含め，すでに高密度マイクロアレイや高速（大規模並列）シークエンサーを用いた解析が始まっている．

3）DNAメチローム（エピゲノム：epigenomeの一部）

メチル化シトシンは5番目の塩基とも呼ばれる．CpG islandあるいはCpG island shoreと呼ばれるCG配列のシトシンのメチル化が遺伝子発現の抑制に重要であると考えられている．DNA配列には変化がないが，細胞機能に大きな影響を与えることから，Epigenetics（エピジェネティクス）/Epigenomics（エピゲノミクス）の研究分野である．解析方法としては，1．メチル化シトシン結合タンパク質特異的な抗体を用いて免疫沈降することでメチル化DNA領域を濃縮し高密度マイクロアレイとハイブリダイズする方法，あるいはメチル化シトシンがbisulfite処理に抵抗性であることを利用して行う，2．高速（大規模並列）シークエンス，3．高密度マイクロアレイ，4．質量分析　などがある．とくに3．高密度マイクロアレイを用いる方法には，Methylation-sensitive single nucleotide primer extensionを応用してメチル化シトシンと非メチル化シトシンを区別するInfinum解析があり，さらに高密度化が進みプローブ数が増加すれば，コスト面からも今後の主流となる可能性がある．包括的DNAメチル化パターンは，疾患の治療反応性や予後予測としてのプロファイルデータとして有用であるし，transcriptomeと相互参照することが重要である．

2．核酸＋タンパク質

ヒストン修飾（エピゲノム：epigenomeの一部）

核酸とタンパク質の相互作用は遺伝子発現制御や，iPS細胞などの細胞リプログラミング技術の観点からも重要視され，染色体の転写活性/不活性領域の解析方法が開発されている．こちらも，DNAメチロームと同様にDNA配列には変化がないことから，エピジェネティクスの研究分野である．この分野で汎用されるクロマチン免疫沈降法（chromatin immunoprecipitation：ChIP）というのは一般的な免疫沈降法をクロマチンに応用したものである．たとえば，ある転写因子が細胞核内で特定のどのようなクロマチン—DNA領域に結合するかどうかを調べたいとする．その細胞を回収しホルマリン処理すると，転写因子とクロマチン—DNAはクロスリンクされる．クロマチンは非常に長いDNAがヒストンとともに凝集している構造なので，これをDNaseや超音波処理により数百塩基程度の大きさにまで断片化する．この断片化された転写因子・クロマチン—DNA複合体に対して，目的の転写因子の抗体を作用させることで，転写因子とクロマチンDNAの複合体を免疫沈降し特異的に濃縮できる．この後，加熱などによってタンパク質とDNAを解離させ，このDNAがどのような領域なのかを調べることができる．調べたいDNA領域に対する特異的プライマーを使って定量的PCRを行うのが，ChIP-qPCRである．この方法では，一度に調べることのできるゲノム領域はごく一部である．PCRに代わり，沈降したDNAを蛍光標識し，高密度マイクロアレイで網羅的に検出する方法があり，これをChIP-chipあるいはChIP-on-chipと呼ぶ．高速（大規模並列）シークエンサーを使って沈降DNAを網羅的に同定する方法もあり，これをChIP-seqと呼ぶ．近年，クロマチンを構成する主要タンパクであるヒス

トンの特定アミノ酸残基の化学修飾（メチル化，アセチル化，リン酸化等）がエピジェネティックな遺伝子発現制御に大変重要な働きを有することが明らかとなっている．これら個別のヒストン修飾を区別できる特異的抗体が市販されているため，特定のヒストン修飾の分布をゲノムワイドに調べることが可能である．ChIP-chip や ChIP-seq は網羅性が高く，これらの技術を使って多くの知見が報告されているが，解析に多くの細胞を必要としたり，少なくとも30種類はあるとされるヒストンの化学修飾を区別して網羅的あるいは経時的に調べるためには，コスト面でも解析技術においてもさらなる改良が必要とされる．

3．タンパク質

プロテオミクス（proteomics）

タンパク質の網羅的解析であるが，タンパク質コード遺伝子の多義性（スプライシングなど）と翻訳後の化学修飾が多様であることで，mRNA と比べて格段に多様性に富み，その機能状態を予測するのが難しい．その数は10万種類以上はあると考えられている．解析方法として２次元ゲル電気泳動による方法が採られることもある．現在はタンパク質の性質や細胞内局在ごとに粗分画し，最初にトリプシン処理を行ってペプチドに分解したものを質量分析計によって網羅的に解析することが多くなりつつある．網羅性という点で，現在はトランスクリプトーム解析に劣り，疾患や薬剤代謝のマーカー探索目的に使用されることが多い．

質量分析による網羅的同定能力は近年発展し（数百から数千種類），さらに網羅的解析にも定量性を与えることが可能となりつつある．定量分析では，炭素や窒素の安定同位体を用いてラベルした既知量のタンパク質（ペプチド）を標品として使い，同位体の分だけ質量荷電比（m/z）がずれたピークの面積を元に計算する方法が主流だが，プロテオミクスは探索的であることが多く，個別の安定同位体ラベルを用意するのは費用面でも現実的ではない．異なる試料間の比を取る（相対定量）のであれば，網羅的解析に対応する方法も開発されている．また，同定されるペプチド数がタンパク質の重量と相関することを利用して，同位体ラベルなしでも定量解析が可能である．これらの技術がより普及すれば，単なる同定を越えて，特異的抗体を使ったウエスタンブロット解析がほぼ不要になるほどの画期的タンパク質同定/定量装置となる可能性があり，経時的データ取得などによって病態形成の理解に貢献できる可能性を秘めている．

4．低分子代謝物

メタボロミクス（metabolomics）

狭義でメタボロミクス（metabolomics）とは TCA 回路，解糖系等の代謝系で生成，分解される中間代謝物や，アミノ酸，脂肪，核酸の代謝中間物など，生化学分野でいう「代謝マップ」に載る比較的低分子の生体内化合物を網羅的に解析する意味で使われる．あらゆる分子が日々入れ替わり代謝されているから，代謝物の omics という名称はそもそも不適切な名称なのかもしれないが，慣習的に，DNA，RNA，タンパク質，糖鎖を除く低分子代謝物に対して用いられることが多い．アラキドン酸などの脂質も含まれることがある．代謝物の特徴として，季節変動，日内変動，食事による変動がよく知られていることから，データの取得はダイナミズム（とくに時間軸）を念頭に行われるべきである．解析対象となる物質特性に応じた質量分析計を組み合わせることで網羅的解析が可能で，診断のためのバイオマーカー探索などに有用である．実際，製薬企業や診断薬メーカーによって，前臨床試験・臨床に利用可能なバイオマーカー探索がメタボロミクスにより行われている．植物の代謝物は動物と比べ桁違いに大量に存在することがわかっており，メタボローム解析による薬剤利用可能な有用天然物探索に期待がかけられている．

5．糖鎖

グライコミクス（glycomics）

糖はエネルギー源としてだけでなく，糖鎖としてタンパク質，脂質に結合し，機能を修飾する重要な役割を持つ．細胞表面にはレクチン型受容体が存在し，細菌などの異物認識機構に関わるし，内在性の免疫反応にも関係していると考えられている．

糖鎖を構成する主要な単糖であるグルコース，ガラクトース，マンノースは，OH 基の配向が異なる分子量180の異性体である．単糖の多糖化に必要なグルコース結合は５つの OH 基のいずれでも結合可能であり，核酸，タンパク質と比べて，構成要素当たりの構造異性体数は桁違いに多い．糖鎖異性体はそれぞれ異なる生物反応あるいは酵素反応に関わることから異性体を見分けることが肝要である．その

ため，質量分析を２回以上行うMS_n法が試みられ，異性体を判別可能な質量分析ピークパタンを得られることがわかってきている．150種類以上ある糖転移酵素を使って人工的に作成した生体糖鎖標準物を元にして質量分析ピークのデータベースが作成されつつある．グライコミクスは発展中の手法であるが，他の解析手法と同じく，疾患診断マーカーや治療標的探索などに期待がかけられている．

データ解釈

個別の遺伝子機能を明らかにしていく伝統的な研究スタイルは今後も重要であることは先に述べた．細胞にはタンパク質，核酸，脂質，糖鎖などの多様な分子がゾル状に詰まって複雑な挙動を取るはずだが，伝統的アプローチによってごく限られた生物現象と分子をリンクすることができる．遺伝子変異と表現型の関係である．網羅的データは生物を一度に大量に情報化する方法にほかならず，「オッカムの剃刀」を信奉するならほとんどの網羅的情報は不要となる．実際の研究では，偏見なく取得した大量の情報からいかに無駄な情報を削ぎ落とし，意味ある情報を抽出するのかが重要視されていて，ごく還元主義的なわれわれが準拠してよい妥当な枠組みである．医学・生物学分野で研究を進める伝統的な研究者は，ほとんどの場合，網羅的データを処理する技術に疎いから，還元的思考を実現できる網羅的データの処理方法が最も必要とされる．

Semantic（意味）解析など

各遺伝子産物の特徴を表すためにGene Ontology Term（GO term）が各遺伝子の付随情報として割り当てられている．Ontologyとは哲学用語で存在論という意味である．GO termは哲学からは離れて，実践的知識利用のためのツールとして作成されている．具体的には，GO termは遺伝子産物の細胞内局在，生物学的機能，分子機能を表す用語からなり，大きな概念的まとまりから，より特異的な機能へと階層性を持たせた単語鎖を，各遺伝子産物に前もって与えたものになる．網羅的データを取り，ある条件で発現が増加するようなリストが抽出できたとき，そのリスト中に濃縮されているGO termがわかれば，生物学的意味を大づかみで解釈可能となるわけである．しかし，そのGO termの質を保証するためには，やはり個別の研究者による遺伝子の定義づけを前提としたいが，現実は違う．生物学的な機能がわかっていない遺伝子がかなりの割合ある．さらに，GO termのデータベースは複数あり，その質は結果にそのまま影響する．GO termに採用される用語についても追加や訂正が必要であろう．GO termを利用したセマンティック解析には有用性と陥穽（データベースの質）が同居するが，網羅的データの解釈ツールとしては誰にもわかりやすく，解析手法としてもシンプルである．

セントラルドグマ（図１）を考えると，タンパク質が細胞機能の主役である．現在はmRNA発現量がタンパク量に近似することを前提に，トランスクリプトームデータから生物学的に意味あるデータを取得しようとしている研究例が多いが，タンパク質量とmRNA量はそれぞれの分解経路が異なったり，翻訳制御機構の存在のために，存在量は一対一対応するとは限らない．論理的にはプロテオームデータが最も適切かつ必須である．mRNAの挙動を使ったネットワーク解析，主成分解析，階層的クラスタリングなどが行われる場合，その解釈には一定の留保が必要である．また，トランスクリプトームデータはその性質から転写制御解析に利用可能であるが，転写制御機構はDNA結合型転写因子だけでなく，DNAメチル化，ヒストン化学修飾などによって複合的に制御されている．現状では複合的な解析には労力と費用がかかるだけでなく，高品質な転写因子結合部位の公的データベースが利用可能でないために，転写制御関係の推定は現状では困難なことが多い．

多層的オミックス解析の可能性

ポストゲノム解析として，RNA，タンパク質，脂質，糖鎖，代謝物等の網羅的データ取得方法が急速に開発され，利用可能となりつつある．今後は技術の発展をうまく捉えながら，同一の試料から複数の解析手法を組み合わせた多層的オミックスデータを取得し，連携させる仕組みが必要であり，本邦での取り組みはまだ始まったばかりである．２種類以上のオミックス解析技術を使った薬用植物（漢方薬）などの解析で，大型プロジェクトが，中国，米，加などでスタートしている（創薬のための有機化合物ライブラリーの多様性が十分でないとの認識から，多様な代謝物を産生する植物から新規な構造をもつ

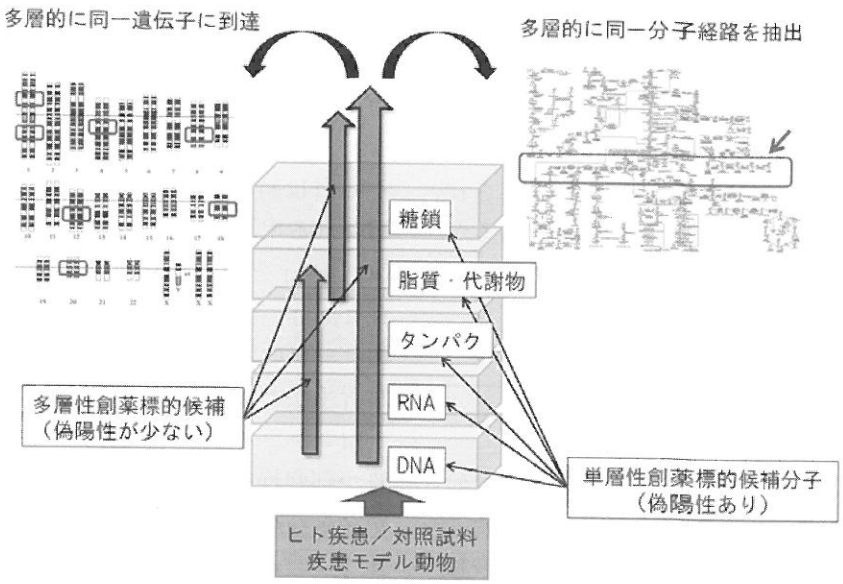

図3　多層的疾患オミックス解析のイメージ

単層性のオミックス解析データのみから疾患原因を見いだすのは困難を極める．そこで疾患由来の試料と対照試料それぞれの多層的なオミックス解析から共通部分を抽出することによって偽陽性を極力抑制し，疾患原因やバイオマーカー同定の効率化を志向する．これらのデータが公開されることで創薬研究，学術研究に役立つと期待される．

天然物を探索することを目的としている）．生物由来のデータはノイズが多いことから，多層的オミックスデータを標準的に用いることにより，意味あるデータを効率よく入手できるだけでなく，疾患あるいは生物現象の成り立ちをこれまでよりも正確に深く理解できることが期待される（図3）．生体分子すべての網羅的データ解析（バイオモレキュロミクス）を推進することによって，疾患の標的分子の発見にも大きな進展が期待されている．そのためには，生体の特徴である時間軸，空間軸を考慮した試料解析，さらなる分析技術の向上，解析ソフトウエアの開発，データベース整備，そしていまだに謎の多い個別の遺伝子産物や生体構成物質の機能解明への「伝統的な取り組み」，「モデル生物での解析」も引き続き必須であり，これらすべてが多層的オミックス解析の成功を担保する．「モデル生物での解析」の意味するところは，生物種間の共通性に注目したヒト以外の生物研究（酵母からサルまで）との「比較」が，結局はヒトを理解するのに効率がよいというところにある．「ヒトはマウスではない」という当たり前の事実が医学研究者からも官庁からも批判される背景には，マウス「モデル」の基礎的研究成果が人間でも完全に同じであるかのように喧伝あるいは誤解される傾向を暗に示している．モデルはモデルに過ぎない．しかしながら，モデル生物研究なしに疾患研究，遺伝子研究は進展しないのも事実である．冒頭で述べたように，タンパク質コード型遺伝子に関する知見だけを取ってみても，ヒトだけを研究してよいほどわれわれの知識は充実していない．

網羅的データを取得するための装置のうち，次世代高速DNAシークエンサーはランニングコストを含めて非常に高額な装置であり，通常の研究室予算の数年分以上に相当するため，日本の非営利機関では限定された施設でしか利用可能でない．他の網羅的データ解析装置の利用についても，合理的かつ積極的政策による推進が必要である．さらに，生物学における網羅的データの最終的解釈あるいは「メタ判断」は計算ではなく人間が行う．多層的オミックス解析にはそのような解釈のできる「医学・生物学者」が必須であり，本邦でもそのための拠点形成，人材確保・育成が今後重要性を増す．

［文献］

1）Lander ES, Linton LM, Birren B et al：Initial sequencing and analysis of the human genome. Nature 2001；409：860-921

2) Takahashi K, Yamanaka S : Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006 ; 126 : 663–76.
3) Cox JS, Shamu CE, Walter P : Transcriptional induction of genes encoding endoplasmic reticulum resident proteins requires a transmembrane protein kinase. Cell. 1993 ; 73 : 1197–206.
4) Mori K, Ma W, Gething MJ, Sambrook J : A transmembrane protein with a cdc2+/CDC28–related kinase activity is required for signaling from the ER to the nucleus. Cell. 1993 ; 74 : 743–56.
5) Murphy WJ, Eizirik E, Johnson WE et al : Molecular phylogenetics and the origins of placental mammals. Nature 2001 ; 409 : 614–8.
6) International Mouse Knockout Consortium, Collins FS, Rossant J et al : A mouse for all reasons. Cell 2007 ; 128 : 9–13.
7) Mitelman F, Johansson B, Mertens F : The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007 ; 7 : 233–45.
8) Quintäs–Cardama A, Cortes J : Molecular biology of bcr–abl1–positive chronic myeloid leukemia. Blood. 2009 ; 113 : 1619–30.
9) Guttman M, Amit I, Garber M et al : Chromatin signature reveals over a thousand highly conserved large non–coding RNAs in mammals. Nature. 2009 ; 458 : 223–7

Exploratory Study of Diseases by Multi–Omics Data

Keisuke Oboki, Kenji Matsumoto and Hirohisa Saito

Abstract Remarkable innovation has been achieved in the fields of comprehensive (omics) data acquisition and analysis. Access to these technologies will allow us to obtain various biological information at a different level from what could be done previously. This review refers to of 1. the innovative technology for analysis of various biological elements. 2. fundamental problems of these omics data. and then. 3. the potential of multi–omics analysis for biology and medicine in the next decades.