バイオデータベースとバイオインフォマティクスツールを知り，学び，使いこなし，研究成果をつかむために

小野浩雅

はじめに

生命科学・医学研究の現場において，バイオデータベースとバイオインフォマティクスツール（DBとツール）は研究インフラとして欠かせない役割を担っている．数多あるDBとツールを取捨選択したり，それらを的確に組合わせて活用する技術の習熟度合いが，日々の研究活動の効率的な進展に影響するだけでなく，研究プロジェクト全体の可能性を左右するといっても過言ではない．一方で，生命科学・医学分野の実験手法や技術革新は目覚ましく，またインターネット技術の驚異的な進歩とともに，その種類や対象，データ量は拡大の一途をたどっている．

DBとツールの種類や対象が日々増加するなかで，初学者にとっては，どのような場面で活用できるのか，まずは何を選ぶべきか，どう使い，組合わせることができるのか，などについて，その全体像を俯瞰し客観的に取捨選択することが難しくなってきている．また，DBとツールの活用についてその重要性を認識しているような経験者であっても，最初に教わったDBとツールを使い続けてしまいがちで，新たなそれらに手を伸ばしづらいという声も多い．

本増刊号は，読者が多種多様なDBとツールを知り，学び，使いこなすため，そして研究という大海原を進んでいくための「羅針盤」となるべく企画したものである．各分野で目的別に厳選した最新のDBとツールを，生命科学・医学を専門とする広いレベルの研究者に向けて紹介している．特にバイオインフォマティクスに明るくない実験系研究者であっても今すぐ使えることを重視し，基本的にコーディングを必要とせずウェブ上で操作が完結するDBとツールをカタログ的に閲覧できることを志向した．また，初学者がステップアップするための足がかりになる情報として，中級者，上級者が着目しているポイントや情報を随所に盛り込むことをめざした．具体的には，検索結果の解釈の方法や考え方の参考になる情報，取捨選択の判断基準，組合わせて使うと有用なDBやツールの紹介について各項目でとり上げている．紹介するDBとツールは研究分野や目的別に全8章，70を数えた．

以下に本増刊号を構成する8つの章の内容について概説する．

第1章．研究を効率化する汎用ツール

第1章では，専門分野を問わず，生命科学研究を効率的に進めるうえで有用なDBとツールについてまとめた．まずは，本増刊号のメインテーマでもある，自分の研究に合うDBとツールを知り，学習するためのサービス（第1章-1～3）について紹介した．また，実験データのもつ意味を解釈するために可視化や統計解析することは日常的に行われているが，それらを手軽に実行することのできるツール（第1章-4～6）も増えてきている．さらに，データの共有や利活用を促進するオープンサイエンス政策が世界的に進められてきていることや，研究の再現性を担保するために，データ解析環境（第1章-7）やプロトコール（第1章-8），図表（第1章-9）の公開・共有サービスを利用することが論文の投稿規定に示されているケースも増えてきている．そして，バイオリソース（第1章-10）は研究開発の効率的な推進のための不可欠な基盤であり，それらの効果的な活用方法について知っておく必要があるだろう．

第2章．文献を調べる・整理する・論文執筆を支援する

第2章では，研究者が日常的に行っている文献の調査・整理や論文執筆について，その効率化を支援するDB・ツールについてまとめた．PubMedを知らない・使わない生命科学研究者は存在しないと思われるが，その利用方法は研究者によって千差万別なのではないだろうか．2020年のアップデートでインターフェースも刷新されたが，今さら聞けないPubMedの使い方（第2章-1）を紹介した．文献や研究成果を効率的に発見し収集するためには，PubMed以外にも有用なツールがあり，日本語の論文や図書・雑誌などの学術情報を検索できるCiNii（第2章-2）はその1つである．PubMedやCiNiiなどを使って得た多くの文献は目的に応じて整理する必要が出てくるだろうし，論文執筆の際には引用文献のリストを作成する必要がある．これらの作業はツールの力を借りることで省力化することができるが，それらの代表的な3つのツール（第2章-3～5）について紹介した．実際に論文を執筆しはじめると，英語表現や関連語（第2章-6, 10），略語（第2章-7），文献間の引用関係の調査（第2章-8）などで思いがけず時間をとられてしまう場合があるが，それぞれについてかゆいところに手が届く有用なツールが提供されている．近年，原著論文の投稿前に草稿（プレプリント）をホストするプレプリントサーバーがさまざまな分野で登場し，新しい知見の迅速な共有とフィードバックを得られる場として活用されている．bioRxiv（第2章-9）は，生命科学分野における代表的なプレプリントサーバーであり，その使い方の実際について紹介した．そして，読者や聴衆の理解を助けるために論文を含むさまざまな研究発表の場で視覚的な図や表を作成することが多いが，いざそれを自分でつくろうとすると時間的・技術的な困難に直面することもある．そのようなときには，利用条件が明記されたイラスト共有サイトや作画ツール（第2章-11）を上手に活用することで効果的

な研究発表を行うことができるだろう．

第３章．ゲノム・遺伝子・NGS データを調べる

第３章では，生命科学研究の基盤として重要なゲノムデータやそれに紐づく遺伝子の配列，機能情報（アノテーション）を効率的に検索し活用する方法についてまとめた．ゲノム配列をはじめとした遺伝情報を生物種ごとにまとめたデータベースはさまざまな研究機関や研究コミュニティによって開発・公開されている．UCSC Genome Browser（第３章−2）やEnsembl Genome Browser（第３章−3）はゲノムブラウザとして著名であり，NCBI Genome Data Viewer（第３章−1）も最近のリニューアルで見やすく整理されている．植物（第３章−4）や微生物（第３章−5）のゲノム情報については国内の研究機関で整備が進められている．研究のさまざまな場面で，ある遺伝子について調べることは多い．多くのDBが遺伝子に関する情報を提供しているが，網羅性や信頼度の点でNCBI RefSeq（第３章−6）がその第一選択肢としてあげられる．このRefSeqの内容をあたかもGoogle検索するかのように，あらゆる検索語で高速に調べることができるGGRNA（第３章−7）というツールも有用である．実験系研究者であっても，一度は配列解析ツールとしてNCBI BLAST（第３章−8）を使用したことはあるだろう．ウェブ版のBLASTだと検索クエリによって検索できなかったり，時間がかかる場合があるが，そのようなときは，GGGenome（第３章−9）を使うと超高速にゲノム配列を検索することができる．また，次世代シークエンサー（NGS）実験によって得られた塩基配列データは国際協力の下で運営されているDB（第３章−10）に日々集積されている．一方，これらの配列データは処理されていないため，そのままではシークエンス結果の解釈ができないが，一部のツール（第３章−11）では，適切なデータ処理を行ったデータを提供している．

第４章．ゲノム編集実験を支援する

第４章では，TALENやCRISPR-Cas9などに代表されるゲノム編集技術を用いた実験をサポートするDBやツールについてまとめた．ゲノム編集による変異導入によって遺伝子を破壊するノックアウト，あるいは任意の配列をある領域へ挿入するノックインなどが行えるが，必ずしも完全ではなく，実験者が意図していない標的配列を切断するケースもあり，これをオフターゲットとよぶ．このオフターゲット作用を考慮してCRISPR-Cas9標的を検索できるツールとしてCRISPRdirect（第４章−1）がある．また同様のツールとして，CRISPROR（第４章−2）があるが，CRISPR-Cas9標的検索だけでなく，そのオフターゲットの検索からプライマー設計までも含めた包括的な情報収集ができる．CRISPR-Cas9やTALENといったゲノム編集ツールではDNA 2本鎖切断（DSB）を引き起こし，その修復の結果発生するフレームシフト変異によってノックアウトが達成されるが，このDSBによって引き起こされる一塩基挿入・欠失変異パ

ターンを予測するInDelphi（第４章−3）というツールもある．また，より応用的なゲノム編集のためのDBやツールもあり，逆転写酵素を利用したゲノム編集技術であるPrime Editingをデザインするための設計支援ソフトウェアPrimeDesign（第４章−4）や，TALENやCRISPR/Cas9などのゲノム編集ツールによって生成された細胞集団内の小さな挿入欠失（indel）の種類と頻度をサンガーシークエンスデータを用いて推定する変異分析解析法・ツールであるTIDE（第４章−5），アンプリコンシークエンシングデータを使った変異分析ツールであるCRISPResso2（第４章−6）についてとり上げている．

第５章．タンパク質やその立体構造について調べる

第5章では，タンパク質の配列や機能，立体構造に関するDBやツールについてまとめた．UniProt（第５章−1）は，タンパク質に関連するさまざまな情報を横断的・網羅的に調べることができる世界で最も広範なタンパク質の情報カタログである．また，プロテオームデータは，細胞内で発現しているタンパク質の完全なセットであり，主に質量分析によってタンパク質プロファイルにもとづいたさまざまな試料の特徴付けが行われているが，jPOST（第５章−2）はその代表的なDBの1つである．一方，Protein Data Bank（PDB）は，タンパク質や核酸，糖鎖などの構造に関する世界唯一の基盤データベースであり，日米欧で形成するコンソーシアムが運営を担当し，国際分業体制で1つのデータとして収集・編集されているが，PDBj（第５章−3）はその日本拠点である．PDBjが提供する立体構造データを表示するための分子ビューアとしてMolMil（第５章−4）がある．また同様に，HOMCOS（第５章−5）は，PDBに記録されているタンパク質複合体などの立体構造データから相同な分子を探索し，構造未知の分子ペアの構造を予測するサーバーである．さらに，2020年にAlphaFold2とよばれるAIシステムによって立体構造予測の分野にブレークスルーがもたらされたが，AlphaFold DB（第５章−6）には，UniProtのヒトのタンパク質と主要なモデル生物のタンパク質の予測立体構造が登録されており，自由に利用することができる．

第６章．実験の結果を解釈する

第6章では，誰でも自由に利用可能な大規模なオミクスデータについて発現解析や機能解析を行い生物学的に解釈するために有用なDBやツールについてまとめた．これらの解析に先立って，生命科学分野のDBのID変換は日常的に行われる作業だが，TogoID（第６章−1）を使うと，幅広いカテゴリーのIDを対象に生物学的な意味付けを参照しながらID変換することができる．マイクロアレイやNGSで測定された発現データは公共レポジトリに登録されるが，NCBI GEO（第６章−2）はその代表的なDBである．欧州のレポジトリであるArrayExpressに登録されたデータは，Expression Atlas（第６章−3）を利用することで，統一的なプロトコールで

再解析された発現データとして比較解析することができる．ある実験条件で得られた発現変動遺伝子リストを機能解析する際には，それらの遺伝子群にどのような機能的な偏りがあるのかを調べるエンリッチメント解析ツール（第６章-4～6）がよく用いられる．同様に，代謝経路やシグナル伝達経路などの分子間相互作用に関するパスウェイデータベース（第６章-7, 8）の情報を利用する場合も多い．近年の著しいウェブ技術の進歩によって，RNA-seqデータ解析はいまやウェブブラウザで実行できるようになっており，機能別に特色のあるいくつかの有用なツール（第６章-9）を利用することができる．

第７章．化合物・代謝産物・糖鎖を調べる

第7章では，質量分析技術や装置の革新によって医療や製薬，化学薬品などのさまざまな研究開発分野で多くのデータが集積されている化合物や代謝産物，糖鎖のDBについてまとめた．PubChem（第７章-1）は化合物に関するDBで，化学構造情報に加えて，どのような目的で行った実験でどのようなデータを取得したかの情報，類似する化合物などが関連づけられている．ChEMBL（第７章-2）もまた，創薬を目的とした生理活性をもつ化合物や小分子のDBで，化合物の情報を部分構造検索や類似性検索で調査することができる．ヒトの代謝産物（メタボローム）に関するメタデータを豊富に収載したDBとしてHuman Metabolome Database（HMDB）（第７章-3）があげられる．網羅的な代謝産物の発現情報を含むメタボロームデータの可視化やその生物学的解釈にはMetaboAnalyst（第７章-4）がよく用いられる．GlyCosmos Portal（第７章-5）は，糖鎖科学のポータルサイトとして，糖鎖情報の登録から，糖鎖に関連する遺伝子，タンパク質，脂質，疾患，パスウェイなどのオミクスデータを網羅的に統合しアクセスすることができる．

第８章．疾患に関するゲノムやバリアントを調べる

第8章では，疾患に関連するゲノムやバリアント情報に関して整理されたDBやよく利用されるツールについてまとめた．dbSNP（第８章-1）は，ヒトゲノムの一塩基多型および短鎖欠失・挿入多型に関するさまざまな情報が登録された網羅的なDBである．gnomAD（第８章-2）は，さまざまな大規模シークエンスプロジェクトから得られたエキソームおよびゲノムシークエンスデータを集約したデータベースで，さまざまなヒト集団におけるゲノムバリエーションとその頻度情報を得たり，構造バリエーションなどの情報を視覚的に得ることもできる．遺伝子型と表現型を比較しながらバリアントの病原性を解釈するためには，ClinVar（第８章-3）がしばしば用いられる．また，染色体不均衡と疾患との関係を収載しているDBであるDECIPHER（第８章-4）では，正確なゲノム位置と過去に報告されている表現型を関連づけることにより，特に人間の発達に影響を与える希少なバリアントについての情報を体系的に理解することがで

きる．全ゲノム関連解析〔Genome-Wide Association Study（GWAS)〕のデータを集積したDBとしてよく用いられるGWAS Catalog（**第８章-5**）では，公開されたGWASの論文からキュレーターが手動で遺伝子型（ジェノタイプ）と表現型（フェノタイプ）の関連を入力し，蓄積したデータを自由に利用可能な情報源として提供している．PolyPhen-2（**第８章-6**）のように，バリアントの機能予測を行うツールも数多く開発されている．多階層オミクス統合DBであるDBKERO（**第８章-7**）では，疾患ヒトゲノム変異の生物学的機能注釈を直感的な操作で行うことができる．

おわりに

本増刊号では，生命科学研究を支える有用なDBやツールについて参考になる利用者が最大公約数的に多いであろうという，編者の独断と偏見にもとづいて，そのごく一部について紹介した．ご自身の関連する分野のDBやツールの稿をご覧いただき試行錯誤するのもよいだろうし，また教科書的に通読して今まで手が伸びていなかった新たな分野への足がかりとする，という使い方もできるだろう．多くの研究者にとっては，これらのDBやツールは，顕微鏡や実験試薬などと同じ「道具」である．便利な「道具」を知って，その使い方がわかれば，あとはそれを活用した研究者自身の情報分析力と想像力の有無が勝負のわかれ目になるだろう．「道具」の使い方如何で，これまで正面からしかみられなかったものが横や後ろやナナメから見ることができてはじめて気がつくことがあるかもしれない．本増刊号を手がかりに，仮説構築からはじまり，実験計画・検証，データ解析，そして論文執筆（以下ループ）という研究サイクルを効率化し加速させることができたのなら，編者としてこれにまさる喜びはない．最後に，本増刊号に執筆を賜った先生方，また企画・編集にあたりお世話になった羊土社「実験医学」編集部の皆さまに厚く御礼を申し上げる．

<著者プロフィール>
小野浩雅：日本大学大学院生物資源科学研究科（応用生命科学専攻・加野浩一郎 教授）に在籍中の2005年頃より，脂肪細胞などの脱分化機構を網羅的に解析するためバイオインフォマティクスを学ぶ．'07年よりDBCLSにリサーチアシスタントとして勤め，特任技術専門員を経て'12年より特任助教．データベースやウェブツールの使い方を動画で紹介する「統合TV」の制作・編集を担当するほか，「RefEx」，「TogoID」，「TogoDX」などの開発に携わる．email：hono@dbcls.rois.ac.jp