

公共データベースを活用した感染症研究：ウイルス探索の一例

堀江真行^{1,2,†}

¹大阪公立大学 獣医学研究科獣医微生物学教室

²大阪公立大学 大阪国際感染症研究センター

[†]責任著者 堀江真行

Tel: 072-463-5694

Email: mhorie@omu.ac.jp

[要約]

ハイスループットシーケンス (HTS) は大量の塩基配列を短時間で取得できる革新的な技術であり、ウイルス探索にも応用され、数々の新規ウイルスが HTS によって発見されてきた。HTS を利用したウイルス探索は感度が高く効率的であるが、そのコストが課題となっており、HTS によるウイルス探索を際限なく行うことは現実的ではない。一方、公共データベースには様々な動物に由来する大量の HTS データが蓄積されている。これらのデータの多くはウイルス探索を目的として取得されたものではないが、ウイルス感染個体より得られたデータであれば、ウイルス由来の配列も含むと考えられる。筆者らはこれまでに、公共の HTS データを用いた大規模ウイルス探索を行い、牛や山羊などの家畜動物を含む多数の動物より新規ウイルスを同定した。さらに、発見したウイルスの一部については、HTS データに付随するメタデータを用いて、標的組織や病原性などに関する詳細な性状解析を行った。これらの研究により、データの再利用によるウイルス探索と解析の有用性が示された。本稿では、公共データベースを利活用した感染症研究の一例として、上記のウイルス探索について概説する。

キーワード：ウイルス探索、ハイスループットシーケンス、次世代シーケンス、データベース、メタデータ

ウイルス探索の意義

地球上には数多くのウイルスが存在するが、脊椎動物のウイルスに限ってもその全容は未だ十分に解明されていない。このような未知のウイルスが今後、人や様々な動物における新興感染症を引き起こす可能性が挙げられている。厄介なことに、ウイルスの中には自然宿主で感染症を引き起こさない、あるいは軽微な感染症を起こす程度である一方で、異種間伝播により他の生物種に重篤な感染症を引き起こすものが存

在する。さらに、他の微生物によって引き起こされる感染症と比較すると、ウイルス感染症は異なる特性を持つ。その一つが感染拡大のスピードであり、ウイルスは他の病原体と比較して非常に速く感染が拡大する。また、費用対効果が高い抗ウイルス薬なども存在しないため、ひとたび拡大しはじめるとウイルス感染症の制御は困難である。ヒトにおいては、上記のウイルスの性質が人獣共通感染症という形で顕在化しており、実際にヒトで起こった多くの新興感染症の多くは、発生当時には未発見の動物由来のウイルスによって引き起こされてきた [11]。さらに、この問題はヒトに限らず、家畜などの動物においても新興ウイルス感染症が報告され

受付：2023年7月28日

受理：2023年8月2日

ている [6]。

以上のような状況から、感染症対策の一環としてのウイルスの探索が改めて注目を浴びている。事前にウイルスの性質や保有動物を知り、感染経路などを予測することによって新興感染症の発生を未然に防ぐとともに、新興感染症が発生した際には、適切な対策を講じることが可能となるであろう。

様々なウイルス探索法とハイスループットシーケンシング技術による変革

古くから様々な生物においてウイルス探索が行われてきたが、その対象としてヒトやヒトと直接関連のある動植物（家畜、伴侶動物、作物など）で顕著な病気を起こすウイルスの探索に偏重が置かれてきたため、それら以外の明らかな感染症と直接的に関連のないウイルスについては十分に研究されてこなかった。後述の通り、現在は幅広い動物からウイルスの探索が行われており、すさまじい速度で新規のウイルスが発見されている。しかし依然として地球上には未発見のウイルスが多数存在すると考えられており、脊椎動物に限っても新規ウイルスに関する報告が後を絶たない。

現在使われている主なウイルス探索法として、ウイルス分離、Polymerase Chain Reaction (PCR、コンセンサス PCR を含む)、ハイスループットシーケンシング (High-throughput sequencing: HTS) が挙げられる。ここではそれぞれの長短所をあげていく。

ウイルスの分離培養：検体を動物や培養細胞に接種し、分離培養を行う。古くから行われている方法であり、今もなおその重要性はゆるがない。他の手法と比べて、増殖可能なウイルスそのものが得られるため、その後ウイルスを用いた解析を直接行うことができるという点において大きなアドバンテージがある。一方で、労力がかかるうえに、専用の施設が必要となる。さらに、熟練の技術を要することも多い。また、該当するウイルスが増殖できる実験動物や培養細胞が必ずしも存在するとは限らないため、全てのウイルスを分離培養できるとは限らない。

PCR：核酸を検体から抽出し、必要に応じて逆

転写を行い、PCRによってウイルス核酸を検出する。ウイルスの核酸を検出するため、培養できないウイルスについても検出可能という利点がある。また技術的にも比較的簡便である。異なるウイルス間で共通した配列にプライマーを設計するコンセンサス PCR であれば、遺伝的に幅広いウイルスを検出可能である。さらには縮重プライマーを用いることにより、より広範なウイルスに対応することができる。現在もある程度標的を絞った検出・スクリーニングではその優位性はゆるがない。一方で、プライマーが必要であるため、既知のウイルスの配列情報に強く依存するという欠点がある。

HTS：次世代シーケンス (Next generation sequencing: NGS) とも呼ばれるが、ここでは HTS と呼ぶ。HTS では検体より抽出した核酸を断片化し、ライブラリを作成して網羅的なシーケンス解析を行う。現在はイルミナ社や MGI 社のプラットフォームがよく利用される。HTS は (深度にもよるが) 存在する核酸を網羅的に検出することができるため、検体にウイルス由来の核酸が含まれる場合には感度よく検出することができる。通常、HTS ではショートリードとよばれる 50 ~ 300 bp 程度の配列断片が得られる。そのショートリードをアセンブルし、得られたコンティグを用いて BLAST などの配列類似性検索を行い、ウイルス配列を検出する (図 1)。HTS によるウイルス検出では PCR のようなプライマーを必要としないため、既知のウイルスの配列情報に依存することなく、非常に幅広いウイルスを検出することができる。なお、厳密にはウイルス様コンティグを検出する際の配列類似性検索において既知のウイルスの配列情報を利用するため、完全に既知のウイルスの配列に依存しないというわけではない。しかし、アミノ酸配列における配列類似性検索では通常、遺伝的にかなり遠いウイルスも検出できるため、多様なウイルスを検出することができる。また、感染ウイルスの量やシーケンスの深度にもよるが、ウイルスゲノム配列を直接決定することもできるという利点もある。HTS は極めて強力なウイルス探索法であり、実際にこれまでに膨大な数のウイルスが HTS によって発見されてきた。家畜感染症領

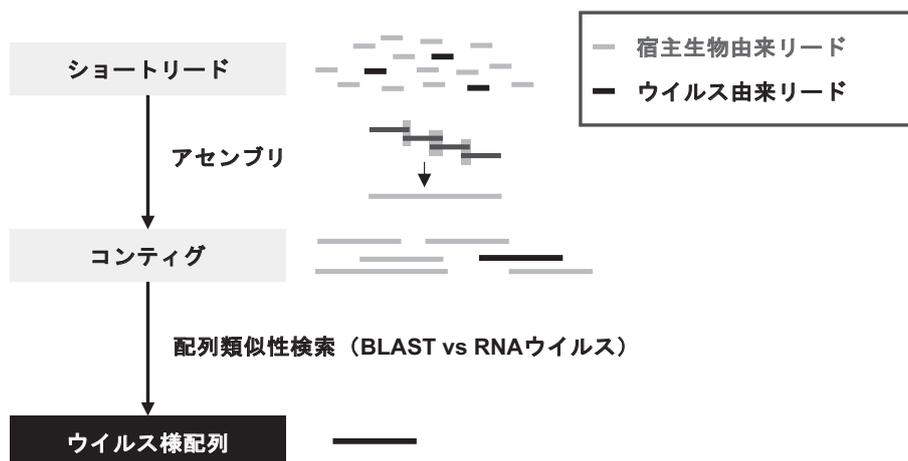


図1 ハイスループットシーケンスデータを用いたウイルス探索法の概要

ハイスループットシーケンスデータを用いたウイルス探索法の概要を記す。ショートリードをアセンブルし、得られたコンティグをクエリーとした配列類似性検索により、ウイルス様配列を検出する。実際は、最後に様々な観点よりウイルス様配列を解析し、ウイルス由来の配列であることを確認する。

域や人獣共通感染症領域においてもその力を発揮しており、ウイルス感染症が疑われる事例やウイルス分離後にウイルスの同定を行うためにも頻繁に利用される。実際に、家畜においてもシマレンベルクウイルス [5]、豚のコロナウイルス [20] など数多くの病原ウイルスがHTSによって同定されている。また、一見健康そうな多数の生物からも大量のウイルスが見ついている。特に近年、中国のグループから報告された大規模なウイルス探索では、既存の知識やウイルスの分類体系を大きく変化させるような多数の新規RNAウイルスが様々な脊椎動物から検出された [18]。一方でHTSによるウイルス探索には欠点も存在する。何より問題となるのがシーケンスにかかる費用である。十年ほど前と比べると驚くほど費用が安くなっているものの、依然として他の手法と比べると費用がかかる。大規模探索では体系的にサンプルをプールすることによって、効率よく探索する方法もとられている [18]。また、検体には宿主細胞由来の核酸が大量に存在するため、たとえウイルスに感染していた検体であっても、得られたリードのうちウイルス由来のリードがごく僅かであることが多い。そのため、検体を核酸分解酵素で処理することによって宿主由来核酸を分解し、ウイルス由来核酸を相対的に濃縮する手法なども取られる [13]。しかし、

これらの工夫をこらしても、依然として大量の検体をハイスループットシーケンスで処理するには膨大な費用がかかるため、ウイルス探索のためだけに際限なくHTS解析を続けることは困難である。

公共データベースを用いたウイルス探索

上述の通り、シーケンスにかかる費用を考えるとHTSによるウイルス探索には限界がある。しかし、世の中にはウイルス探索に供されていない大量のHTSデータが存在する。公共データベースに登録されているHTSデータである。DDBJ DRAやNCBI SRAなどといった公共データベース [1] には、大量のHTSデータが存在する。HTSは多岐にわたる生命科学分野で活用されており、多くの学術雑誌において、論文で使用したHTSデータの公共データベースへの登録が義務付けられている。これらの公共データベースには当然ながら様々な動物に由来するデータも大量に存在する。このようなデータの多くは、ウイルス探索以外の研究において取得されたデータであるが、その由来となる動物がウイルスに感染していた場合には、ウイルス由来の配列も含まれると考えられる。つまり、公共データベースには未知のウイルス由来の配列が眠っている可能性があり、それらのデータを「再利用」することによって、シー

クエンスコストをかけずに効率よくウイルスを探索できるかもしれないということである。筆者らはこのような公共データに着目し、様々な動物由来の HTS データの解析を行った [7-10]。その一例を紹介する。

筆者らは合計 46,000 以上の哺乳類、鳥類由来の RNA シークエンス (RNA-seq) データを網羅的に解析し、RNA ウイルスの検出を試みた (図 2) [9, 10]。本研究においては、より計算コストを下げるために、検体の由来となる生物種のゲノム、あるいは由来となる生物種と同じ属の種のゲノム配列が公開されている場合は、そのゲノム配列にショートリードマッピングし、マップされなかったリードのみを用いることによって、宿主由来のリードを除去した。その後アセンブルを行い、得られたコンティグ配列を用いた配列類似性検索によって RNA ウイルスの配列を検出した。その結果、上記の 46,000 以上の RNA-seq データより、少なくとも 22 の科にわたる約 900 個の RNA ウイルスの感染を検出することができた。

筆者らはこれらのウイルスのうち、少なくと

も 9 個の新規ウイルスについてほぼ全長のゲノム配列を決定することができた。その中には畜産動物である牛のパレコウイルス (bovine parechovirus) や山羊のヘパトウイルス (goat hepatovirus) も含まれる (図 3)。

牛パレコウイルスはピコルナウイルス科パレコウイルス属のウイルスである。人のパレコウイルスにおいては小児の種々の疾患に関連があるとされている [2]。様々な動物からパレコウイルスが検出されているが、家畜動物からの検出は初の報告であった。また、筆者らが発見した牛パレコウイルスは遺伝的に新規なパレコウイルスであり、国際ウイルス分類委員会 (International Committee on Taxonomy of Viruses: ICTV) によって定められたパレコウイルス属の新種の基準を満たしていた。後述の通り (「メタデータを利用した解析」の項を参照)、現在は疾病との関連は不明であるが、日本国内においても牛パレコウイルス感染牛が存在することが Nagai らによって示されている [15, 16]。

山羊ヘパトウイルスはピコルナウイルス科へ

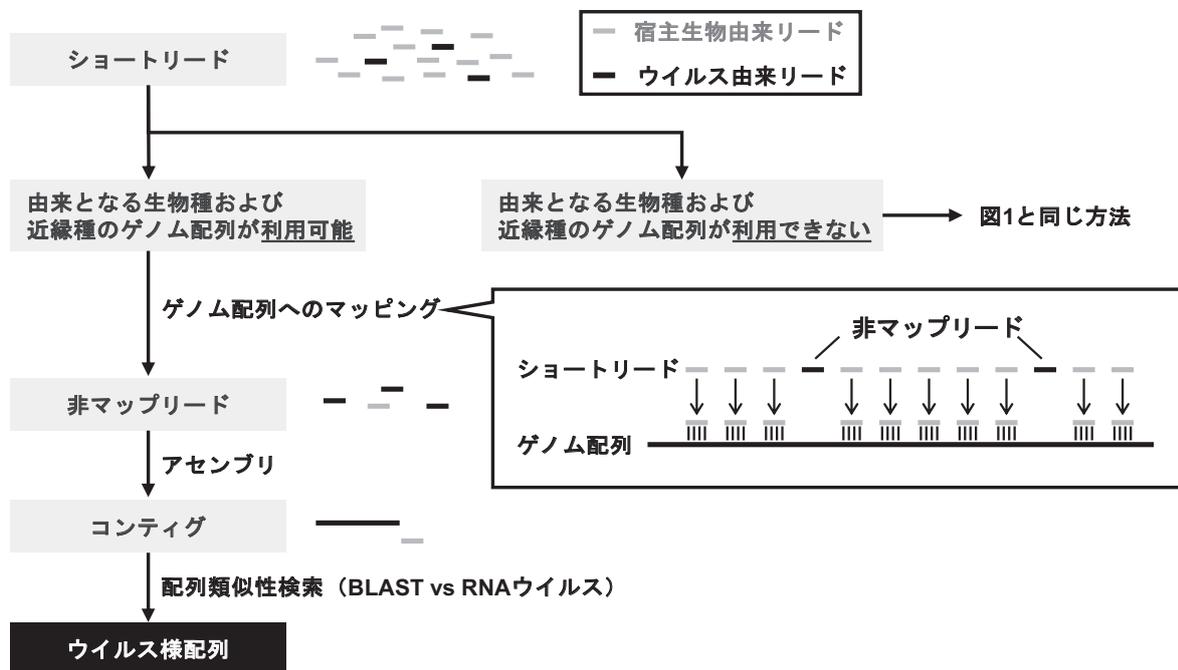


図2 筆者らが行った公共データベースを利用したウイルス探索法の概要

筆者らが行った公共データベースを利用したウイルス探索法の概要を記す。データの由来となる生物種、あるいはその近縁種のゲノム配列が利用可能な場合にはマッピングを行い、マップされなかったリード (非マップリード) を用いてアセンブルする。これらをスーパーコンピュータ上で並列して行うことにより、大量のデータを効率よく解析することができる。

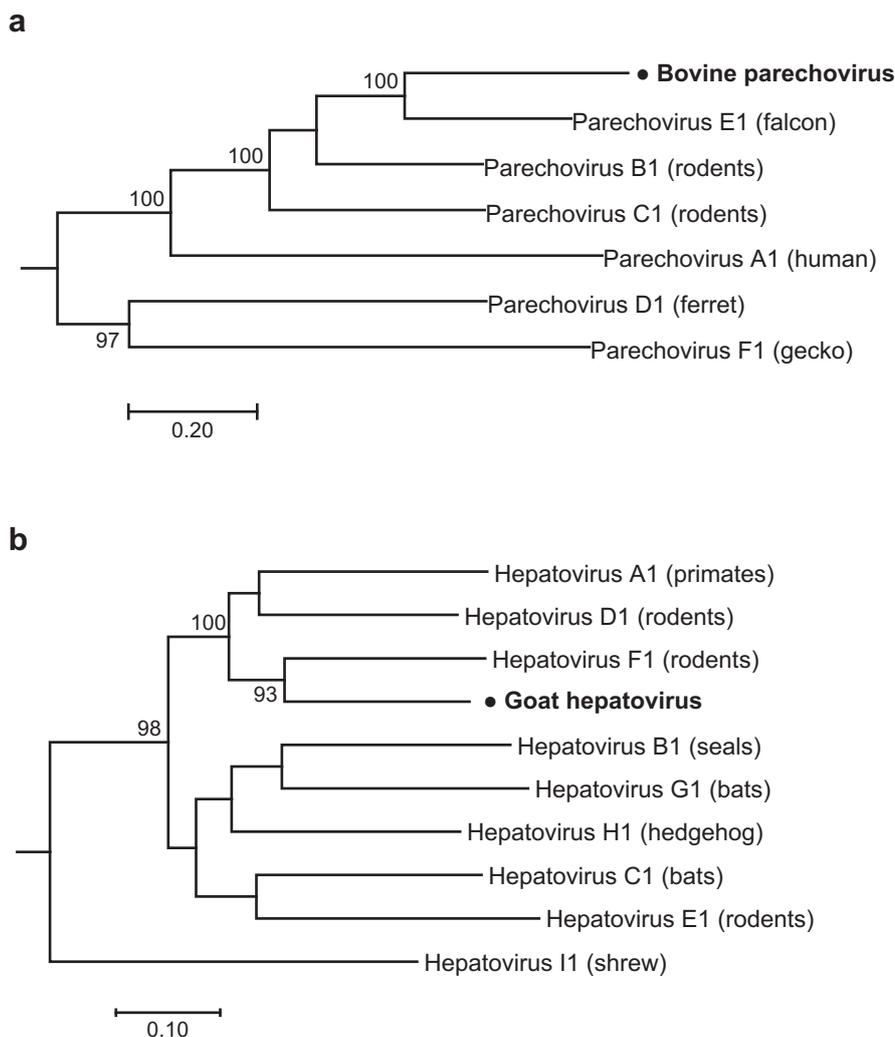


図3 牛パレコウイルスと山羊ヘパトウイルスの分子系統樹

牛パレコウイルスと山羊ヘパトウイルスの分子系統樹を示す。筆者らの論文のデータを用いて再解析した。枝上の数字はブートストラップ値を示す。70未満のブートストラップ値はここでは削除している。スケールバーは遺伝距離（置換/部位）を示す。筆者らの論文 [10] のデータを再解析した。

パトウイルス属のウイルスである。ヘパトウイルス属の代表的なウイルスとして、ヒトの病原体である A 型肝炎ウイルスが知られている。ヘパトウイルスも様々な動物から検出されているが、家畜動物の検出は報告されておらず、初めての報告となる。山羊ヘパトウイルスはげっ歯類のヘパトウイルスとともに A 型肝炎ウイルスとクラスターを形成していた。

その他にも、様々な畜産動物から新規のウイルスが検出されており、筆者らは現在も詳細な解析を行っている。このように筆者らはデータを再利用することにより、効率的にウイルス探索を行うことが可能であることを実証した。

マッピングによる大規模ウイルス検出

公共データベースを用いたウイルス研究は、新規ウイルスの探索のみならず、発見したウイルスがどのような動物に感染しているのか、大規模な解析をするのにも役立つ。上記のショートリードのアセンブルと配列類似性検索を組み合わせ合わせた手法（図 1、2）は感度が高いものの、計算コストが高く膨大な時間がかかるため、特定のウイルスを標的としてそのウイルスの感染を大規模に検出することには適していない。一方、リードを標的ウイルスの配列にマッピングすることによって、比較的少ない計算コストで

効率よくウイルス由来のリード (= ウイルス感染) を検出することができる。さらに、マップされたリード数を数えることによってウイルス核酸を半定量することができる (図 4)。

筆者らは実際に発見したいいくつかのウイルスについて、マッピングによる大規模解析を行った。後述の通り、このマッピングにより HTS データからウイルス由来のリードを検出し、さらにはそのデータに付随するメタデータの解析を組み合わせることによって、ウイルスについて様々な情報が得られる。マッピングによる特定の標的ウイルスに対する解析は計算コストもそれほど高くないため、比較的容易に解析できる。新規ウイルスに限らず、もし興味の対象となるウイルスがあれば、マッピングによる大規模解析を行うことによって様々な知見が得られるかもしれない。さらに、宿主遺伝子発現変動解析などを組み合わせることにより、新しいウイルス宿主間相互作用が見えてくるかもしれない。

メタデータを利用した解析

公共データベースの配列データには通常メタデータが付随しており、配列データを取得した検体の由来となる生物種、組織 (臓器や細胞)、健康状態、地域、さらには検体の採取方法やライブラリの作成法、シーケンス法など、様々な情報が記載されていることが多い (下記の通

り、これらの情報は登録者の労力に依存しているため、必ずしもすべてが存在するわけではない)。これらの情報を用いることによって、検出したウイルスについてさらに詳細な情報を得ることができる。いくつか実例を紹介する。

上記の通り、筆者らはデータ解析によって牛パレコウイルスを発見した。マッピングによる解析の結果、牛パレコウイルスは非常に多くの牛由来の HTS データより検出された。幸いにも牛パレコウイルスが検出されたデータに付随するメタデータとして比較的詳細な情報が記載されていたため、それらを利用することによって、標的臓器や疾病との関連についての解析を行うことができた。

まず、個々の HTS データについて、牛パレコウイルスゲノムにマッピングをされたリード数をカウントするとともに、検体の由来となる組織の情報を組み合わせることによって、牛パレコウイルスの標的臓器を調べた。その結果、中枢神経系、消化器、リンパ組織など幅広い組織から牛パレコウイルス由来のリードが多く検出され、牛パレコウイルスは広範な組織に分布することが示唆された (図 5a)。さらには一部のメタデータには牛の健康状態 (健康、消化器疾患、呼吸器疾患) についても記載があったため、各種疾病との関連性について疫学的な解析を行った。しかし、これらの疾患とパレコウイルス感染については有意な関連は見られなかつ

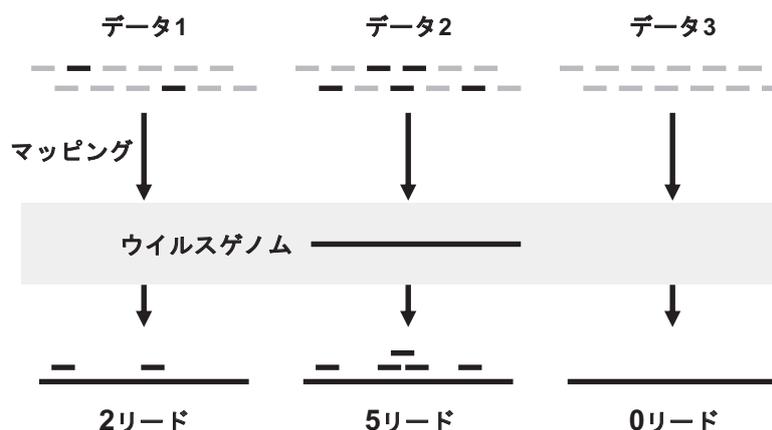


図4 マッピングによるウイルス由来リードの検出

マッピングによるウイルス由来リードの検出とカウントの概要を示す。標的となるウイルスゲノムにマッピングを行い、マッピングされたリード数をカウントすることによってウイルス感染を検出するとともに、ウイルス由来核酸量の半定量が可能となる。

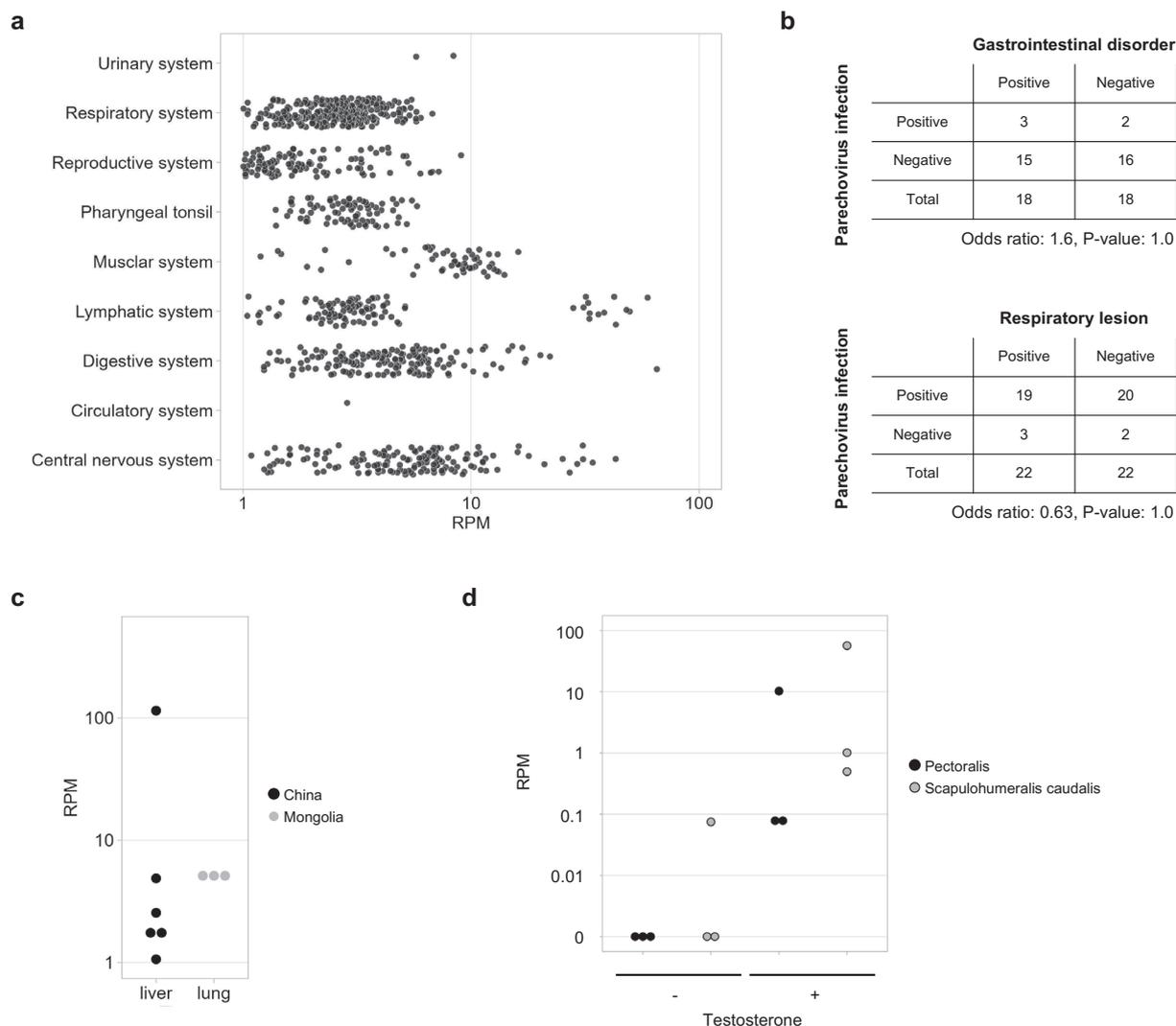


図5. マッピングとメタデータの組み合わせによる解析の例

マッピングとメタデータの解析を組み合わせることにより、ウイルスの様々な性状を解析することができる。(a) 各臓器における牛パレコウイルス由来のリード数の分布を示す。RPMはread per millionを示す。(b) パレコウイルスと臨床症状についての解析を示す。(c) 肝臓と肺における山羊ヘパトウイルス由来のリード数と、データが取得された国を示す。(d) テストステロン処理の有無による鳥デルタウイルス由来リード数を示す。2つの組織に由来するリード数を個別に表示している。筆者らの論文 [9, 10] のデータを再解析した。

た (図 5b)。

山羊のヘパトウイルスについても同様の解析を行った結果、感染する臓器や分布域についての情報が得られた。山羊ヘパトウイルスはモンゴル、中国由来の複数の個体から検出され、少なくとも東アジアの山羊においてある程度感染が広がっていることが示唆された (図 5c)。また、山羊ヘパトウイルスはヘパトウイルスの本来の標的組織である肝臓に加え、肺からも比較的多くのウイルス由来リードが検出された (図

5c)。山羊ヘパトウイルスについても病原性は不明であるが、上記の通り A 型肝炎ウイルスに比較的近縁であるため、家畜の感染症はもちろんのこと、人獣共通感染症という観点からも注意が必要かもしれない。

また、ウイルスと宿主の相互作用についても解析することができる。私たちは以前、様々なスズメ目の鳥より新規デルタウイルスを発見した [9]。マッピング解析の結果、この鳥のデルタウイルスはスズメ目の幅広い鳥種より検出さ

れ、世界中に分布することがわかった。さらに、興味深いことに鳥のデルタウイルスがテストステロンの投与実験に用いられた数羽の鳥より検出された。これらの鳥では、テストステロン非投与群と比べ、投与群においてはるかに多いウイルス由来リードが検出された（図5d）。テストステロンは鳥においても免疫抑制作用が報告されており [3]、おそらくテストステロン処理によって免疫が抑制され、それによりウイルスの複製が活発になったのであろう。すなわち、この鳥のデルタウイルスは宿主生物の免疫反応に感受性が高いのかもしれない。

このように、公共データを用いることによって効率よくウイルスを探索できるだけでなく、配列解析とメタデータの解析を組み合わせることによって、ウイルスについてさらなる情報を得ることができる。上記の通り、既知のウイルスについてもこのような解析によって、新しい現象を発見できるかもしれない。

公共データベースを用いたウイルス探索の欠点

ここまで、公共データベースを用いた解析の良い点ばかりを挙げてきたが、実際にはいくつかの欠点や、解析・解釈における注意点がある。まずは配列の「コンタミ」である。サンプリング、核酸抽出、ライブラリ作成、さらにはシーケンスの際のインデックスホッピング [12, 14, 19] など、本来検体に含まれるはずのない配列が最終的に得られるデータに混入することがある。実際に筆者らの公共データの解析においても、コンタミとしか考えられない多数の配列（A型インフルエンザウイルスの研究室株である A/WSN/1933 (H1N1) やレトロウイルスベクターなどに利用される水疱性口内炎ウイルスの G 遺伝子の配列など）が検出され、それらは詳細な解析から除外した [10]。また、実際に筆者自身がライブラリを作成し、シーケンスをした際にも、研究室内の他のメンバーが取り扱っている核酸配列のコンタミを経験したこともある。

データの由来となる検体が手元があれば HTS データで得られた結果を検証できるが、公共データの場合、得られた配列がコンタミであるのか、それとも本当に検体に存在する核酸配列であるのか、明らかなケースを除いて検証

するのは困難である。もちろん、データの登録者に連絡することによって検体にアクセスできることもある。しかし、筆者の経験からすると連絡が取れないことも少なくなく、また連絡が取れた場合においても既に検体が破棄されているなど、検体にアクセスできないことが多い。

検体にアクセスできないことにより、ウイルスゲノム配列の決定の際にも問題が生じることがある。ウイルス由来のリード数が十分でない場合、ウイルスゲノムの（ほぼ）全長のコンティグが得られず、断片的なコンティグしか得られない場合がある。この際、検体にアクセスできる場合は PCR や RACE 法などにより、コンティグにカバーされなかった領域の配列を決定することができる。しかし、検体が利用できない場合、これらの配列については決定することができないため、断片的なゲノム配列情報しか得られない。

さらに HTS に限ったことではないが、由来となる検体が検出されたウイルスの真の宿主であるかどうかは慎重に解釈する必要がある。特に腸管由来の検体では、食餌や常在微生物に感染するウイルスである可能性も否定できない。

メタデータが十分でないというケースも少なくない。筆者らが以前に行った解析 [10] において、一部のデータでは由来となる動物種のみしか記載されておらず、由来となる組織すらわからないこともあった。このような場合、せっかくウイルスを同定しても得られる情報は非常に少なくなってしまふ。このように、公共データを利用した解析における様々な欠点や制約を念頭に置く必要がある。

今後 HTS 解析をされる皆様へのご願い

現在、HTS の価格はそれほど高額ではなくなり、多くの研究者が多様な研究に利用するようになった。それに伴い、公共データベースに登録されるデータの数も飛躍的に増加している。すなわち、ウイルスの探索はもちろんのこと、様々な研究に「再利用」できるデータがますます増えつつあり、データの再利用により様々な仮説の検証やデータ駆動型の研究がしやすくなるということである。一方で、上記の通り、少なくない数の配列データにおいて付随するメタデータが十分でなく、解析に十分に活用

できないことが多い。上記の通り、実際に今回のウイルス探索においても十分な情報が得られないこともあった。そこで本稿をご覧になった皆様へのお願いとなるが、配列データを登録される際には、ぜひ登録できる限りの情報を登録していただけるようお願いしたい。このようなメタデータを含めたデータの「質」が、データを再利用する研究においては極めて重要な要素となる。手間がかかるうえに直接の「見返り」もほとんどないため省略されがちであるが、科学の発展のためにもぜひともお願いしたい。

今後の展望

データの再利用はウイルス探索に限ったことではなく、様々な感染症関連の研究に有用である。鳥のデルタウイルスの例で示した通り、病原体側だけでなく、宿主側も含めて解析することも可能である。筆者らは詳細な解析は行っていないものの、宿主の遺伝子発現解析などを含めた総合的な解析も可能である。現在は個人で使用しているコンピューターのスペックも以前とは比較にならないほど向上しているため、小規模の解析であれば自前のコンピューターでも十分に解析可能である。また、コマンドが必要なキャラクターユーザーインターフェース (CUI) のような一般的に馴染みのない方法だけでなく、グラフィカルユーザーインターフェース (GUI) での様々な解析サービスも充実しつつあるため、それらを利用すれば比較的簡単に解析ができる。一方で、大規模な解析となると膨大な計算資源が必要となるため、手元のコンピューターや GUI での解析では処理しきれない。しかし、国立遺伝学研究所などのスーパーコンピューターは無料で使用することができる。また、筆者が利用している東京大学のヒトゲノム解析センターのスパコンの有料プランも、アカデミックユーザーであれば 4,500 円 / 月でそれなりの大規模な解析が可能である。これらの場合、ある程度の CUI の基礎知識が必要となる。現在では日本語の書籍はもちろんのこと、解説ウェブサイトや解説動画なども充実しており、筆者が解析に取り組み始めた頃と比べると随分と解析の敷居が低い。各自が公共のデータを有効に活用することによって、仮説の検証やデータ駆動型の研究などを効率よく進め

ることが可能な時代である。

なお、筆者が公共データベースをウイルス探索に利用できると気がついたのは、2010 年頃に全く別の目的で公共データベースの cDNA ライブラリの配列を用いた検索をしていた時に偶然、一見健康と思われるジュウシマツ由来の cDNA ライブラリからボルナウイルスの配列を大量に見つけたときである [17]。このときに筆者はデータを本来の目的とは別の目的に「再利用」することができるという公共データベースの強みに気がつき、データ解析によって新規ウイルスを発見するという着想を得た。筆者はいわゆるバイオインフォマティクスではなく、独学バイオインフォマティクスユーザーとして研究を展開してきた。そのため、筆者らがこれまでに行った解析パイプラインなどが洗練されているとは言い難い。実際に、より洗練された大規模ウイルス探索も報告されている [4]。しかし、筆者が生粋のウイルス学研究者であったからこそ、詳細なデータをウイルス学的に有効活用できたと自負している。このように、今後もバイオインフォマティクスを主軸としない研究者が、公共データを自在に活用することによって、科学が大きく前進すると信じている。

最後に、公共データを利用した解析はあくまでも多々ある研究手法の一つであるということ強調したい。データ解析のみならず、様々な手法による実験やさらなる検証を組み合わせることによってさらに研究は深化するであろう。実際に筆者らは、データ解析によって同定したウイルスの人工合成系を作成し、ウイルス学的な実験を行うことによって様々な性状を明らかにした [9]。今後、公共データの再利用によりさらなる感染症研究が発展することを祈る。

引用文献

- [1] Arita, M., Karsch-Mizrachi, I. and Cochrane, G. 2021. The international nucleotide sequence database collaboration. *Nucleic Acids Res.* 49: D121–D124.
- [2] Britton, P. N., Jones, C. A., Macartney, K. and Cheng, A. C. 2018. Parechovirus: an important emerging infection in young infants. *Med. J. Aust.* 208: 365–369.
- [3] Duffy, D. L., Bentley, G. E., Drazen, D. L. and Ball,

- G. F. 2000. Effects of testosterone on cell-mediated and humoral immunity in non-breeding adult European starlings. *Behav. Ecol.* 11: 654–662.
- [4] Edgar, R. C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B., Banfield, J. F., de la Peña, M., Korobeynikov, A., Chikhi, R. and Babaian, A. 2022. Petabase-scale sequence alignment catalyses viral discovery. *Nature.* 602: 142–147.
- [5] Hoffmann, B., Scheuch, M., Höper, D., Jungblut, R., Holsteg, M., Schirrmeyer, H., Eschbaumer, M., Goller, K. V., Wernike, K., Fischer, M., Breithaupt, A., Mettenleiter, T. C. and Beer, M. 2012. Novel orthobunyavirus in Cattle, Europe, 2011. *Emerg. Infect. Dis.* 18: 469–472.
- [6] Höper, D., Mettenleiter, T. C. and Beer, M. 2016. Metagenomic approaches to identifying infectious agents. *Rev. Sci. Tech.* 35: 83–93.
- [7] Horie, M. 2021. Identification of a novel filovirus in a common lancehead (*Bothrops atrox* (Linnaeus, 1758)). *J. Vet. Med. Sci.* 83: 1485–1488.
- [8] Horie, M., Akashi, H., Kawata, M. and Tomonaga, K. 2021. Identification of a reptile lyssavirus in *Anolis allogus* provided novel insights into lyssavirus evolution. *Virus Genes.* 57: 40–49.
- [9] Iwamoto, M., Shibata, Y., Kawasaki, J., Kojima, S., Li, Y.-T., Iwami, S., Muramatsu, M., Wu, H.-L., Wada, K., Tomonaga, K., Watashi, K. and Horie, M. 2021. Identification of novel avian and mammalian deltaviruses provides new insights into deltavirus evolution. *Virus Evol.* 7: veab003.
- [10] Kawasaki, J., Kojima, S., Tomonaga, K. and Horie, M. 2021. Hidden Viral Sequences in Public Sequencing Data and Warning for Future Emerging Diseases. *MBio.* 12: e0163821.
- [11] Kawasaki, J., Tomonaga, K. and Horie, M. 2023. Large-scale investigation of zoonotic viruses in the era of high-throughput sequencing. *Microbiol. Immunol.* 67: 1–13.
- [12] Kircher, M., Sawyer, S. and Meyer, M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40: e3.
- [13] Kohl, C., Brinkmann, A., Dabrowski, P. W., Radonić, A., Nitsche, A. and Kurth, A. 2015. Protocol for metagenomic virus detection in clinical specimens. *Emerg. Infect. Dis.* 21: 48–57.
- [14] Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L. and Graf, J. 2014. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS One.* 9: e94249.
- [15] Oba, M., Obinata, S., Takemae, H., Kazama, K., Oguro, M., Ito, K., Kakinuma, S., Ishida, H., Murakami, H., Sakaguchi, S., Mizutani, T. and Nagai, M. 2023. Prevalence and genetic diversity in bovine parechovirus infecting Japanese cattle. *Arch. Virol.* 168: 91.
- [16] Oba, M., Sakaguchi, S., Wu, H., Fujioka, Y., Takemae, H., Oki, H., Kawai, M., Shiokawa, M., Aoki, H., Fukase, Y., Madarame, H., Nakano, T., Mizutani, T. and Nagai, M. 2022. First isolation and genomic characterization of bovine parechovirus from faecal samples of cattle in Japan. *J. Gen. Virol.* 103.
- [17] Rubbenstroth, D., Rinder, M., Stein, M., Höper, D., Kaspers, B., Brosinski, K., Horie, M., Schmidt, V., Legler, M., Korb, R. and Staeheli, P. 2013. Avian bornaviruses are widely distributed in canary birds (*Serinus canaria f. domestica*). *Vet. Microbiol.* 165: 287–295.
- [18] Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J., Liu, L., Holmes, E. C. and Zhang, Y.-Z. 2018. The evolutionary history of vertebrate RNA viruses. *Nature.* 556: 197–202.
- [19] Wright, E. S. and Vetsigian, K. H. 2016. Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics.* 17: 876.
- [20] Zhou, P., Fan, H., Lan, T., Yang, X.-L., Shi, W.-F., Zhang, W., Zhu, Y., Zhang, Y.-W., Xie, Q.-M., Mani, S., Zheng, X.-S., Li, B., Li, J.-M., Guo, H., Pei, G.-Q., An, X.-P., Chen, J.-W., Zhou, L., Mai, K.-J., Wu, Z.-X., Li, D., Anderson, D. E., Zhang, L.-B., Li, S.-Y., Mi, Z.-Q., He, T.-T., Cong, F., Guo, P.-J., Huang, R., Luo, Y., Liu, X.-L., Chen, J., Huang, Y., Sun, Q., Zhang, X.-L.-L., Wang, Y.-Y., Xing, S.-Z., Chen, Y.-S., Sun, Y., Li, J., Daszak, P., Wang, L.-F., Shi, Z.-L., Tong, Y.-G. and Ma, J.-Y. 2018. Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. *Nature.* 556: 255–258.

Discovery of viruses using public sequencing data

Masayuki Horie

¹ Laboratory of Veterinary Microbiology, Graduate School of Veterinary Science, Osaka Metropolitan University

² Osaka International Research Center for Infectious Diseases, Osaka Metropolitan University

[Abstract]

High-throughput sequencing (HTS) is an innovative technology that enables the rapid acquisition of large numbers of nucleotide sequences. This technology has been employed in virus detection, leading to the discovery of numerous new viruses. However, the cost of HTS is high, making it impractical to unlimitedly conduct virus discovery using HTS. On the other hand, a large amount of HTS data from various animals has been accumulated in public databases. Although most of the publicly available HTS data were not obtained for virus discovery, some virus-derived sequences may exist if the data were obtained from virus-infected animals. We have previously conducted a large-scale virus search using public HTS data and identified novel viruses from various animals, including domestic animals such as cattle and goats. In addition, we characterized some of the discovered viruses using metadata from the database. These results demonstrate the utility of virus discovery and analysis through data reuse. This paper presents our aforementioned research as an example of infectious disease research using public databases.

Keywords: database, high-throughput sequencing, metadata, next-generation sequencing, virus discovery