

第2回 SPARC Japan セミナー2018

「オープンサイエンス時代のクオリティコントロールを見通す」

生命科学研究におけるプレプリント活用の現状

坊農 秀雅

(情報・システム研究機構 ライフサイエンス統合データベースセンター)

講演要旨



プレプリントの活用が生命科学系においても急速に広まりつつある。プレプリントサーバーは、査読される前の結果だけでなく、これまで情報が共有されにくかった新型の計測機器によるデータの「癖」や、査読誌ではカバーしづらかった公共データベースやウェブツールに関する詳細な記述（ドキュメント）の置き場としても活用されている。2018年に入ってから、演者も生命科学系のプレプリントとしてよく使われるBioRxiv（バイオアーカイブ）に公共データベースからのメタ解析の研究論文を一本アップロードしたほか、別の共同研究の論文も査読誌に投稿する前にアップロードしてある。その現状について実例を交えながら紹介しつつ、いかにして情報の質を担保するかについて、演者の考えるところを議論したい。



坊農 秀雅

理化学研究所においてFANTOM (Functional annotation of mouse) プロジェクトの立ち上げに関わった後、埼玉医科大学ゲノム医学研究センターを経て、2007年7月より情報・システム研究機構 ライフサイエンス統合データベースセンター (DBCLS) にて統合データベースプロジェクト (統合DB) に従事。統合DBの広報・普及活動として統合TVや統合データベース講習会AJACSの立ち上げに関わり、現在は国立遺伝学研究所にて日本DNAデータバンク (DDBJ) と連携して大規模塩基配列データ検索技術の開発とそれを活用した生命科学研究を行っている。京都大学博士 (理学)。

私の所属は情報・システム研究機構のライフサイエンス統合データベースセンターで、静岡県三島にある国立遺伝学研究所のDNA Data Bank of JapanのDNAデータを扱いやすくしたり、実際にそのデータを使った研究もしたりしています。

私はこの1年でプレプリントが身近になったので、これを機会に自分の1年をまとめてみたいと思います。

1~2年でオープンアクセスはどう変わったか

図1は、私が2年前にこのSPARC Japan セミナーで出した、生命科学研究者の立場から見てオープンアクセスはどのようなものかという雑感を書いたものです。この2年で変わったのは、bioRxiv というものが、新しいことを考えている人には知られるようになったと

いうことです。生命科学研究者は、『PLOS ONE』や『Scientific Reports』といったオープンアクセス誌のことは知るようになって、自分もそういうところに出たいと思うようにはなりました。しかし、グリーン

生命科学研究者の立場から見たOpen Access (OA)

- ・ 現状、ゴールド OA一択
 - ・ すなわち「オープンアクセス」=ゴールドOA
- ・ グリーンOAという言葉自体ほとんど知られてない
 - ・ IFの付いてない雑誌は回避される傾向
 - ・ ピアレビューを経た論文のみ業績とみなす伝統

第1回 SPARC Japan セミナー2018 「オープンアクセスへの道」
坊農秀雅「生命科学研究分野における研究者の投稿先雑誌選択傾向とOAへの意味づけ」
スライド一部改変

(図1)

OA という言葉はほとんど知られていない、インパクトファクターの付いていない雑誌は回避される傾向がある、ピアレビューを経た論文のみ業績とみなす伝統があるという状況はあまり変わっていないのではないかと思います。

ちょうど昨年の今ぐらいに、『Science』という雑誌がプレプリントに関する特集を組んで、生命科学でもプレプリントがあり、論文が出されているということを紹介してくれました（図 2）。このスライドも去年出しています。

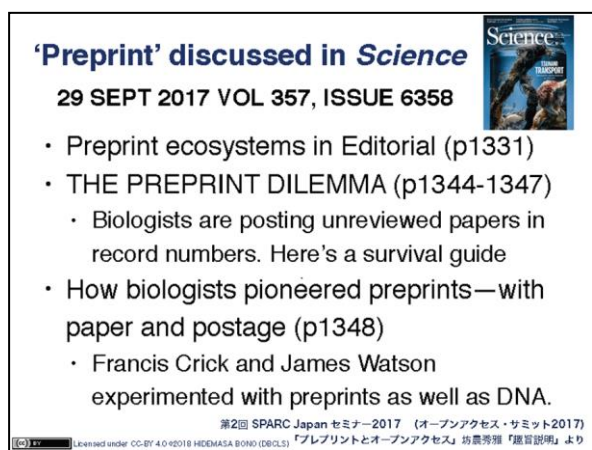
私が bioRxiv フリークになった理由

昨年、ここで 10 月最後の月曜日に SPARC Japan セミナーに出させていただいて、その週の途中から Cold Spring Harbor 研究所に出張で参りました（図 3）。私の研究分野はパイオインフォマティクスという分野

で、生命科学でも普段から実験するのではなく、データを相手にして、DNA のデータから何か新しいものを見つけたり、データベースをつくったりしています。その研究で有名な研究所です。DNA の二重らせんのモデルのブロンズ像が置いてあります。bioRxiv は Cold Spring Harbor Laboratory Press というこの出版会がやっているものです。

そこで学会があったのですが、その学会に行くとこんなものが置いてあって、たくさん T シャツを配っていました（図 4）。今日はその T シャツを着ています。帰ってベッドで広げて写真も撮りました。自分のノートパソコンに bioRxiv のシールを貼っているぐらい、bioRxiv フリークになりました。

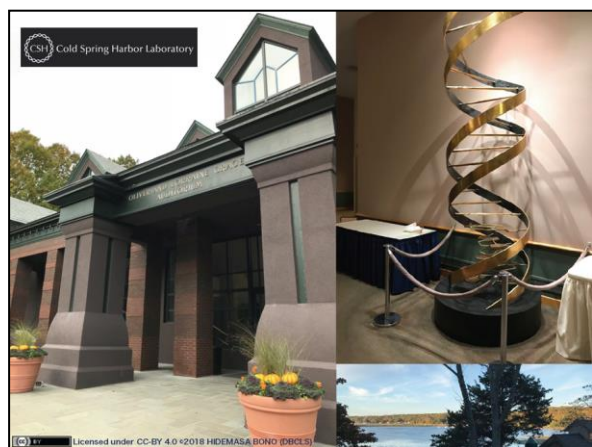
それは理由があって、皆さん、学会で発表するからです。口頭発表とポスター発表があり、口頭発表者はもちろんなのですが、ポスター発表者でも、全部自分



(図 2)



(図 4)



(図 3)



(図 5)

がやっている研究は bioRxiv に論文としてアップロードしているのです。昔だったら、「まだ論文投稿中で、あまり詳しくは話せないのだけれども」と言っていたのが、bioRxiv の ID、つまり DOI をみんな書いてあるのです。それにショックを受けて、「これは自分もやらなければいけない」と思いました。

「次の論文はプレプリントサーバー (bioRxiv) にまず出そう」という決意をして、その決意を証拠に残すために Twitter でもその日につぶやきました。11 月 4 日という、去年の今ぐらいですが、証拠が残っています (図 5)。

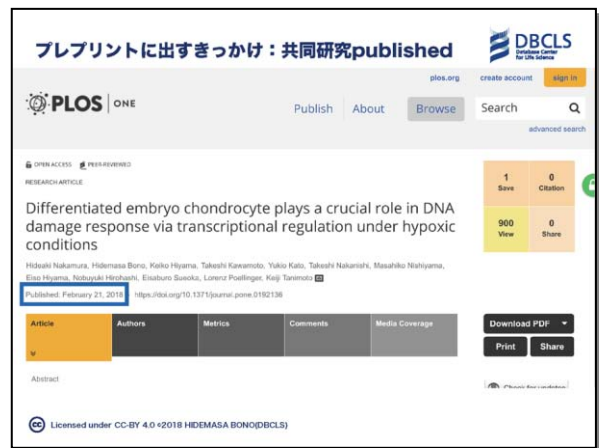
その雑感をさらに Twitter に書いて「月曜日の SPARC Japan セミナー『プレプリントとオープンアクセス』に始まって、水曜日から土曜日まで bioRxiv をホストしている Cold Spring Harbor Laboratory での Genome Informatics Conference。すごくプレプリントな 1 週間だったなあ」と言っていたぐらい、その 1 週間でプレプリントをやりたい、出したいという人になりました。

bioRxiv へのサブミット

帰ってすぐやれたらよかったのですが、いろいろやっていて 2 月になって、出す機会がついにめぐってきました。図 6 は私が出したプレプリントのウェブサイトのスクリーンショットです。「Meta-analysis of hypoxic transcriptomes from public databases (公共データベースの中の低酸素刺激による遺伝子発現データのメタ解

析)」という題名の論文をプレプリントサーバーにアップしました。これはなぜ出したかという、きっかけがありまして、図 7 も実は『PLOS ONE』に出したのですが、私の共同研究の論文が、2018 年 2 月 21 日に出たのです。私自身がデータベースから検索して見つけてきたことをきっかけに、共同研究者が実験をして発見があった、その報告を『PLOS ONE』に出したのです。その私が見つけてきたというのが、ただ単にデータベースを検索したわけではなく、データベースからいろいろデータを集めてきて、バイオインフォマティクスの計算をして、キュレーションをした結果、見つけられました。

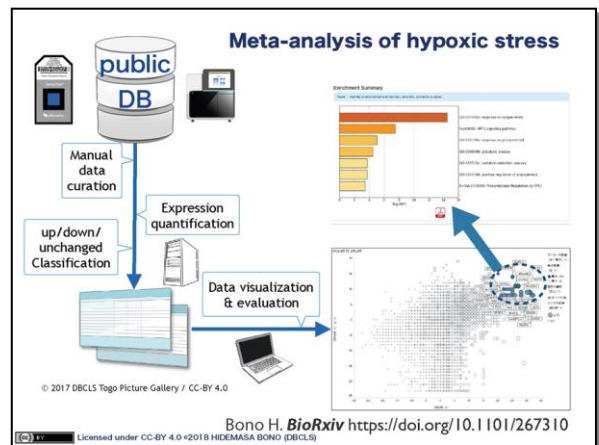
その結果も、ただ単にこれを見つければいいという一個だけの遺伝子ではなく、たくさんある中から見つけてくださいと共同研究者に遺伝子リストを渡してありました。共同研究者はそれを長らく 2 年ぐらい見て



(図 7)



(図 6)



(図 8)

いて、その結果見つけてきて論文になったのです。

私自身がやった手法を、本論文には詳しくは書けなかったもので、それをどこかに残したいからプレプリントに書こうということで、ずっと前から書いてあったのですが、この論文が出版されたので、それに合わせて bioRxiv に投稿しました。これは日にちが大事なのですが、先ほどの論文は 2 月 21 日にパブリッシュされ、私はその日のうちに bioRxiv にサブミットしました。

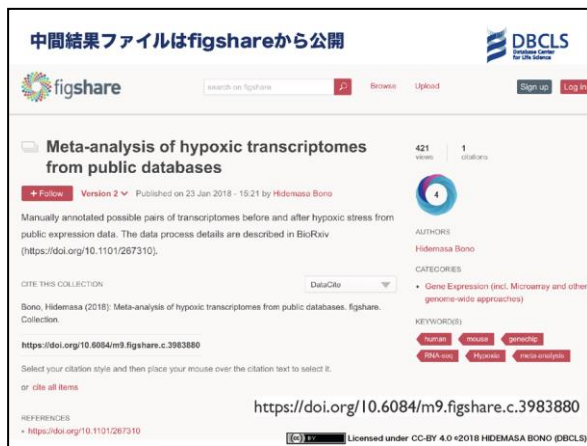
2 月 23 日にこれは間違いなくオリジナルの論文だと認めていただいて、DOI もめでたく付いて、公開されることになりました。この日は静岡県民にとっては大事な日で、富士山（223）の日です。富士山の日にめでたく出せたということです。

この研究は実際に何をやっているのかというのを少しだけ紹介させていただくと、公共データベースを使

っていて、これは遺伝研にある DNA のデータベースなのですが、その DNA のデータベースを使って、人間が酸素がうすいところに行ったらどんなことになるかを調べました。これは人間の実験で、あまりやると死んでしまうので、できないのです。普通の実験で使う酸素濃度はとてもうすいです。私たちがよく使うのは 1%の酸素濃度です。普段皆さんは 21%のところに住んでいますが、1%の酸素濃度になったときにどういふふうに変化するかを見るために、いろいろな人がおこなった実験からデータを取り出して比較しました。

DNA の配列そのものではなく、タンパク質になる前に出てくるメッセンジャーRNA の数がどれぐらい変わるかを調べました（図 8）。

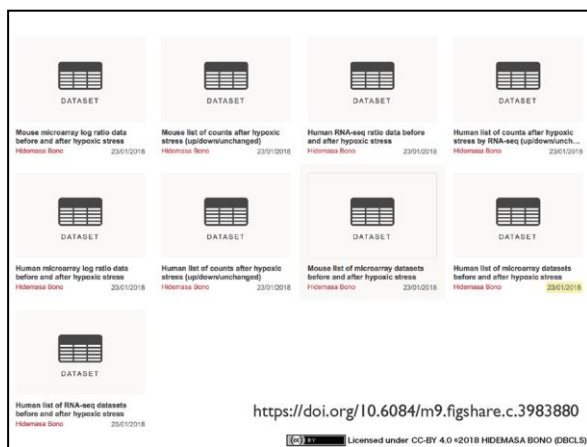
そういうことをおこなったデータがたくさんデータベースに入っているのを、それを計算したデータのフ



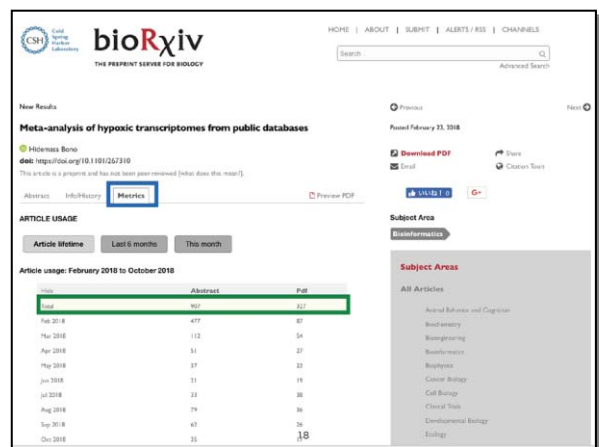
(図 9)



(図 11)



(図 10)



(図 12)

ファイルをつくりました。最終的には、論文にするときは可視化するのですが、実際に大事なものはこのような中間データです（図 9）。このあたりのデータは専門家が見るとよく分かるので、中間データを皆さんでシェアしたい、このデータは再利用しがいがあるということで、bioRxiv に出しました。ただ単に bioRxiv に出すだけだったら普通の成果なのですが、このようなテーブル形式のデータをサブミットしたいということで、テーブル形式の Excel のデータは figshare にアップしました。

figshare にアップすると、同じようにデータテーブルに DOI が付与されます（図 10）。ここからダウンロードできるようにして、この figshare の DOI を先ほどの論文にたくさん書いて登録しました。

データテーブルの中身は図 11 のように、遺伝子の ID と数字がたくさん書いてあります。

bioRxiv の「Metrics」のタブを選択すると、何回見られたかが分かります（図 12）。これは 2 月にアップしたばかりですが、おかげさまでトータルで 900 回ぐらい見られています。論文全文の PDF がダウンロードされた回数も月ごとに出て、よく分かります。これも論文をプレプリントにアップした人間としては励みになります。

実は、こういうことがあるよと私の共同研究者に言うと、共同研究者は私より前に bioRxiv にアップしていたのです。その共同研究者から、一緒に論文をアップしようということで、2018 年 7 月にアップして、3

カ月ぐらいでこれぐらい見られていますということが分かりました（図 13）。自分がメインで出した論文を 1 本、共同研究でもう 1 本、bioRxiv にこの 1 年で出すことができました。

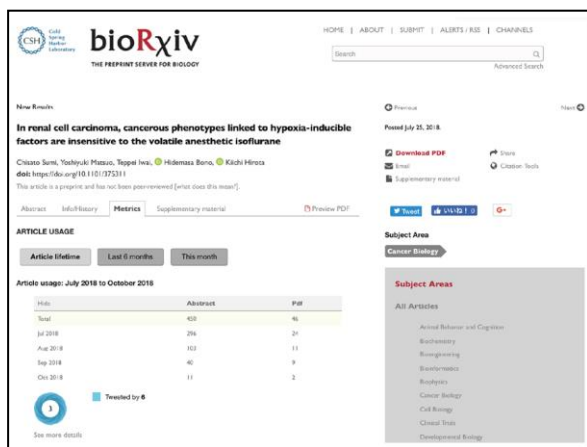
生命科学研究者は、アメリカ国立衛生研究所（NIH）の国立生物工学情報センター（NCBI）が運営している PubMed で論文を検索して、そこに載っていたら、みんなが成果と認めるという暗黙の了解があるのです。PubMed ではまだこのプレプリントは検索できないのですが、Europe PMC では検索できます（図 14）。

Europe PMC をご存じない方もいるかもしれませんが、ヨーロッパでも PubMed 的なアクティビティをしていて、PMC は PubMed Central の略で、アメリカの PubMed の全文版が PMC と呼ばれて、利用可能になっています。PubMed はご存じのとおり抄録、アブストラクトだけなのですが、PMC は全文が見られるコレクションです。これも数が増えています。

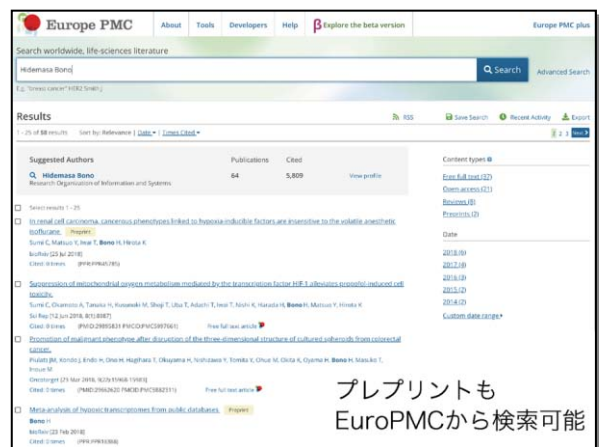
基本的には同じデータがあるのですが、Europe PMC のインターフェースでは、例えば私の名前で検索すると、bioRxiv に登録されたようなプレプリントも一緒に検索できます。このようにプレプリントと分かるように書いてあって、PubMed ID は当然ないのですが、検索できるという状況になっています。

プレプリントのメリット・デメリット

出してみても分かったプレプリントのメリット・デメリットについてお話しします。メリットは、すぐ公開



（図 13）



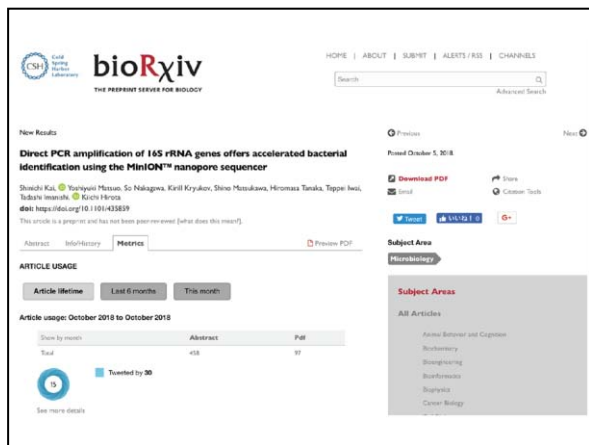
（図 14）

できる、DOI で引用可能である、現状は article processing charge (APC) 無料、査読誌への投稿ができる、Europe PMC から検索可能という点です。特に私がメリットだと思ったのは、DOI で引用可能なために、研究費申請のときにリファレンスとして引用することができるとことです。また、私が勧めた友達が良かったと言っていたのは、先に論文を出せたということです。生命科学者の中では、論文を投稿している間にレビュアーがそれを見て、それをまねして実験を始めて、出し抜くことがあるといわれています。それが阻止できる、先取性が担保できるのがいいと思っています。

現状で感じているデメリットは、この辺はよく理解されていないから起きていることなのですが、一度そのように公開すると別のところに論文が出せないかもしれないという恐怖です。ほとんどの場合は時代が追い付いてきて、そんなことはなくて、多くの雑誌に論文を出すことはできるという意識が広まると思います。

また、今までやってきた査読論文に出すスタイルを変えたくないとか、査読論文しか評価されない現実があるとか、そういう後ろ向きなことですが、このあたりもできれば変えていけたらと思っています。

私の共同研究者は、果敢にも今月も、とても注目されるようなプレプリントを出しています(図 15)。これは査読論文を出す前にこういうことを出しているのですが、今月だけで 500 を超えるぐらい見られていて、こういうものをうまく使うと注目される、という形に



(図 15)

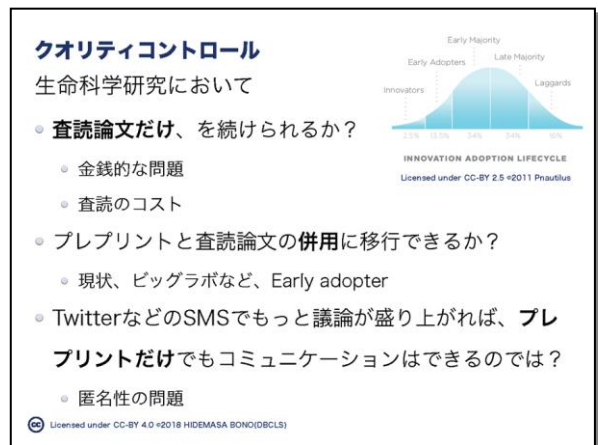
できるのではないかと考えています。

クオリティコントロール

最後にクオリティコントロールの話をしてします。われわれの生命科学研究において、今までやってきた査読論文だけを続けられるかということがあります。査読にはコストがかかります。午前中に、『Scientific Reports』がだいぶ遅くなっているという話がありましたが、それは私たちも感じています。査読していただくということだけでは無理ではないかという実感があります。

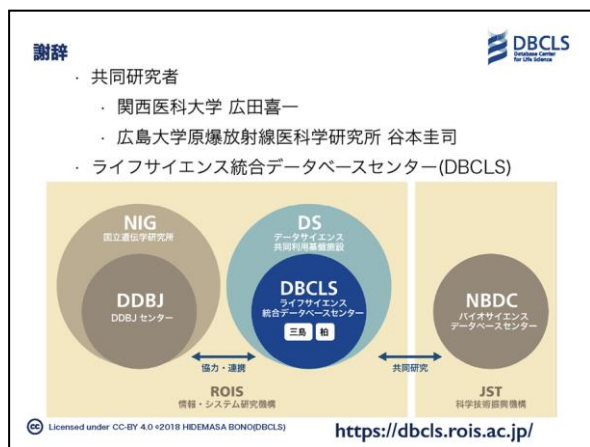
次の段階として、プレプリントと査読論文の併用に移行できないかと考えていますが、プレプリントを出すというのは、図 16 の分布で言うところの Early Adopters ぐらいしかできていないので、まだ様子見のレベルではないかと思います。もっと議論が盛り上がっていけば、プレプリントだけでもっとコミュニケーションができるのではないかと思います。

海外は分からないですが、日本の Twitter をする人は、研究者と名乗っては言っているものの、匿名でやりたい人が多くて、議論は盛り上がっているようなのですが、誰か分からないということが多く、そのあたりが邪魔するのだろうかと考えています。ぜひ実名でこういう議論ができるようになればいいのですが、実名でやってしまうと、自分のプロモーションなどに影響するのではないかと考えている人が多いのではないかと思います。



(図 16)

この研究は、関西医科大学の広田喜一さんと、広島大学の谷本圭司さんという方との共同研究で行ったものです（図 17）。普段こういう研究をする機会を与えていただいているのは、ライフサイエンス統合データベースセンター（DBCLS）です。DBCLS は図の真ん中の組織で、今回の SPARC Japan セミナー企画ワーキンググループメンバーであるバイオサイエンスデータベースセンター（NBDC）の八塚さんとは共同研究を普段させていただいています。この国立情報学研究所も情報・システム研究機構（ROIS）の中の研究所であるということを最後に紹介して終わりにしたいと思います。



(図 17)