„WAS IST DER AUSWEG AUS DEM DILEMMA ZWISCHEN SCHNELLER FREIGABE VON SEQUENZDATEN UND ERSTVERÖFFENTLICHUNGSANSPRUCH? GEHEIMHALTEN HILFT WENIGEN UND SCHADET VIELEN. ES GIBT KEINE ALTERNATIVE ZU EINEM SCHNELLEN FREIEN DATENZUGANG!“

Rudolf Amann

Öffentliche Sequenzdaten sollten rasch wirklich frei zugänglich sein!

DOI: 10.1007/s12268-019-1016-z

Wir leben und forschen in Zeiten von *Big Data*. Die Qualität, mit der wir uns durch diese „schöne neue Welt“ bewegen, hängt von öffentlich zugänglichen Datenbanken ab. Hier sind Informationen nicht nur abgelegt, sondern mit relevanten Metadaten verknüpft. So finden wir in einer fremden Stadt nicht nur unseren Weg, sondern auch ein gutes und offenes Restaurant mit der gewünschten Speisekarte. Für Lebenswissenschaftler sind Datenbanken vor allem Sequenzdatenbanken. Hier legen wir den deskriptiven Teil unserer Forschung häufig in Form von Genomen und Metagenomen ab, hier entwickeln wir unsere Hypothesen und testen sie im Licht der neu hinzukommenden Daten.

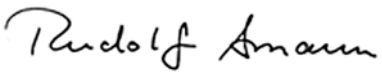
Als Doktorand kannte ich Mitte der 1980er-Jahre noch jedes meiner 10.000 selbst sequenzierten Nukleotide persönlich, bevor dann in den 1990ern die Sanger-Sequenzierung von Tausenden klonierten 16S-rRNA-Genen folgte. Ende der 1990er Jahre wurden die ersten Bakteriengenome entschlüsselt und zehn Jahre später Metagenome von Umweltproben. Heute habe ich keine Bedenken, einer Doktorandin ein kombiniertes Metagenom/Metatranskriptom-Projekt zur Frühjahrsbakterienblüte vor Helgoland vorzuschlagen, das Sequenzanalysen von mehreren 10^{12} Nukleotiden umfasst.

Heute sind Terabasenpaare (10^{12}) – und bald auch Petabasenpaare (10^{15}) – schnell sequenziert. Die Analyse ist nun der zeitaufwendige Teil. Auch können diese gigantischen Sequenzdatensätze nicht mehr umfassend analysiert werden. Entweder betrachtet man sie aus großer Flughöhe oder entlang sehr spezifischer Fragestellungen. Häufig entsteht so eine zeitliche Lücke zwischen öffentlicher Verfügbarkeit der Datensätze und einer ersten Publikation durch die Datenproduzenten. Denn viele öffentliche Sequenzierzentren schalten die Daten innerhalb weniger Monate frei, während die Analyse und Publikation Jahre in Anspruch nimmt. Eine möglichst rasche Datenfreigabe von Sequenzdaten wurde schon 2003 im Fort Lauderdale Agreement gefordert, das aber auch noch ein Recht der Datenproduzenten auf eine erste Publikation formulierte [1]. Das Dilemma ist, dass dann freigegebene Daten nicht wirklich nutzbar sind, da noch keine Veröffentlichung vorliegt. Dieses Problem ist in den letzten Jahren so groß geworden, dass ich mit vielen weiteren Kollegen vorgeschlagen habe, das Anrecht auf eine erste Veröffentlichung abzuschaffen [2]. Daten sind erst dann frei verfügbar, wenn sie von allen für Publikationen genutzt werden können.

Das kann dazu führen, wie es uns am Bremer Max-Planck-Institut passiert ist, dass ein gut aus unseren Nordsee-Metagenomen assemblierbares Genom von australischen Kollegen für eine Veröffentlichung genutzt wurde, bevor wir unsere zu 99,98 Prozent identische Datenanalyse eingereicht hatten. Schade, aber die im Joint Genome Institute in Walnut Creek, Kalifornien, mit dem Geld amerikanischer Steuerzahler produzierten Sequenzen waren ja auch in einem internationalen *Community Sequencing Project* für die Allgemeinheit produziert worden.

Was ist der Ausweg aus dem Dilemma zwischen schneller Freigabe und Erstveröffentlichungsanspruch? Einfaches Geheimhalten hilft wenigen und schadet vielen. Es kann daher keine ernsthafte Alternative sein, obwohl ich die Angst vor „Datenpiraterie“ gut verstehen kann. Denn faktisch ist die Sequenzanalysekapazität zwischen Arbeitsgruppen ungleich verteilt. Zunehmend differenzieren sich die Produktion von Daten und deren Analyse. Dabei geht heute fast die ganze Anerkennung an denjenigen, der die Analysen publiziert. Hinter Planung und langjähriger Beprobung von Langzeitobservatorien oder Patientenkohorten stecken aber nicht nur Arbeit und Logistik, sondern auch wissenschaftliche Konzepte. Auch die DNA- und RNA-Isolierung war für die Datenproduzenten aufwendig, von den Sequenzierkosten, der notwendigen Qualitätskontrolle und der Arbeit des Hochladens in der öffentlichen Datenbank ganz zu schweigen.

Mit meinen Kollegen schlage ich daher vor, eine rasche Freigabe durch eine höhere Anerkennung der Datenproduzenten zu motivieren. Letztlich ist die Ermöglichung des Zugangs – selbst als einfache *accession number* – ein sehr wichtiger Schritt. Jeder Datensatz sollte dafür unbedingt mit einem *Digital Object Identifier* (DOI) verknüpft werden. DOIs sind zitierbar, die Nutzung der Daten kann damit besser nachverfolgt werden, und sie können im Lebenslauf genutzt werden. Zudem könnte eine Datenfreigabe in einer Kurzpublikation mit Autorenliste, Methodenteil und Danksagung bekanntgegeben werden, wenn es dafür spezialisierte Journale geben würde. Erstveröffentlichung und rasche Datenverfügbarkeit wären so gewährleistet. ■



Rudolf Amann, Direktor der Abteilung Molekulare Ökologie, MPI für Marine Mikrobiologie, Bremen

Korrespondenzadresse:
Prof. Dr. Rudolf Amann
Max-Planck-Institut für Marine Mikrobiologie
Celsiusstraße 1
D-28359 Bremen
Tel.: 0421-2028-930
ramann@mpi-bremen.de

Literatur

[1] The Wellcome Trust (2003) Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. Meeting Report, www.sanger.ac.uk/legal/assets/fortlauderdalereport.pdf
[2] Amann RI, Baichoo S, Blencowe BJ et al. (2019) Toward unrestricted use of public genomic data. Science 363:350–352