Fruzsina Molnár-Gábor, Jan Korbel & das GHGA-Konsortium

Das Deutsche Humangenom-Phenomarchiv

Menschliche Genomdaten und verwandte Omics-Daten (z.B. Transkriptome, Proteome) sind inhärente Bestandteile der biomedizinischen Forschung geworden und sind damit Bausteine der zukünftigen Gesundheitsversorgung.[1]

Bestehende Forschungsdateninfrastrukturen können die spezifischen Anforderungen der unionsrechtlichen und deutschen mitgliedstaatlichen Vorgaben für die Datenverarbeitung derzeit nicht ausreichend erfüllen. Darüber hinaus sind die bestehenden Infrastrukturen im engen Sinne als Archiv konzipiert und bieten der Forschungsgemeinde keine effiziente und benutzerfreundliche Plattform für die Datenanalysen und die Replikation von Ergebnissen auf anderen Kohorten. Die bisher fehlende Möglichkeit, menschliche genomische Daten in Deutschland adäquat zu archivieren und gemeinsam für die Forschung zu nutzen, schränkt die Nutzbarkeit dieser Forschungsdatensätze stark ein und stellt insbesondere ein Hindernis für die Verknüpfung und Integration von Daten dar, die bei der Erzielung von aussagekräftigen Forschungsergebnissen im Omics-Bereich erforderlich sind.

Das Deutsche Humangenom-Phenomarchiv (GHGA)[2] wird eine Lösung für diese Herausforderung bieten, indem es eine nationale Infrastruktur für die sichere Speicherung, das Zugriffsmanagement, den Austausch und die Analyse von menschlichen Omics-Daten bereitstellt. Um eine harmonisierte, technisch interoperable Infrastruktur zu schaffen, ist GHGA eng mit existierenden nationalen Omics-Datenlieferanten und deren IT-Infrastrukturen verknüpft. Durch die Nutzung bereits bestehender Infrastrukturen sollen zunächst die Hürden der Datenarchivierung reduziert werden. Zugleich wird GHGA als nationaler Knotenpunkt in die föderierten europäischen und internationalen Dateninfrastrukturen eingebunden, hierbei spielt das *European Genome Phenome Archive* (EGA) eine besondere Rolle.[3] In einem nächsten Schritt soll die Funktion des GHGA über die reine Archivierung hinaus ausgebaut werden, indem über eine cloudbasierte Infrastruktur die vergleichende Analyse von Omics-Datensätzen ermöglicht wird. Rechenintensive Anwendungen innerhalb der Cloud bedeuten eine besondere Herausforderung für die Ausgestaltung dieser Infrastruktur und der Datensicherheit, jedoch ermöglichen solche Strukturen auch eine Verarbeitung der Daten durch Forscher ohne das sensible Daten heruntergeladen werden müssen. Die zentrale Speicherung und der Zugriff über ein cloudbasiertes System können dadurch bei sachgerechter Umsetzung mehr Sicherheit bieten und begünstigen den „Nutzer", da er selbst keine große Rechen- und Speicherkapazitäten zum Auslesen und zur Analyse der Daten benötigt. Des Weiteren sind personalisierte Portale für spezifische Nutzergruppen sowie die Kuratierung von besonders wertvollen und häufig genutzten Referenzdatensätzen geplant. Diese werden die Vorteile der Infrastruktur für die verschiedenen Nutzergruppen aus Grundlagenforschung und klinischer Versorgung steigern. Umgekehrt werden die Nutzer selbst an der Entwicklung der Infrastruktur maßgeblich mitarbeiten. Ebenfalls beteiligt und repräsentiert in der Governance der Infrastruktur sind PatientInnen-Vertreter.

Die GHGA-Infrastruktur wird mit Förderbeginn zum 1. Oktober 2020 im Rahmen der Nationalen Forschungsdateninfrastruktur (NFDI) aufgebaut.[4] Die NFDI setzt sich zum Ziel, die Datenbestände von Wissenschaft und Forschung systematisch und fachübergreifend zu erschließen, nachhaltig zu sichern und zugänglich zu machen sowie (inter-)national zu vernetzen. Sie wird in einem aus der Wissenschaft getriebenen Prozess als vernetzte Struktur eigeninitiativ agierender Konsortien aufgebaut werden. Die Verarbeitung von Daten mit Personenbezug wird als Querschnittsthema zwischen den NFDI-Konsortien angesehen.[5] In der Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen wird zu ihrer Verarbeitung der Bedarf an gemeinsamen Best Practice-Lösungen genannt, mit denen die

Dr. iur. Fruzsina Molnár-Gábor
Heidelberger Akademie der Wissenschaften, BioQuant-Zentrum
Im Neuenheimer Feld 267 (BQ 049), 69120 Heidelberg
Fruzsina.Molnar-Gabor@hadw-bw.de

Dr. Jan Korbel
Europäisches Laboratorium für Molekularbiologie (EMBL)
Meyerhofstraße 1, 69117 Heidelberg, Germany
korbel@embl.de

DOI: .....

[1] Die Omics-Forschung basiert auf bioanalytischen Hochdurchsatzanalysetechniken, mit denen sich molekulare Profile von Zellen, Geweben und Tumoren in kurzer Zeit erstellt lassen. Dadurch können molekulare Veränderungen detailliert erfasst werden. Nationale Akademie der Wissenschaften Leopoldina, 2014: Zukunftsreport Wissenschaft. Lebenswissenschaften im Umbruch – Herausforderungen der Omics-Technologien für Deutschlands Infrastrukturen in Forschung und Lehre. Deutsche Akademie der Naturforscher Leopoldina e.V. Nationale Akademie der Wissenschaften. Halle/Saale.

[2] GHGA, Letter of Intent within the Call for National Research Data Infrastructures (NFDI), S. 6, https://ghga.dkfz.de/documents/Letter_of_Intent_GHGA.pdf. Kontaktaufnahme unter contact@ghga.de.

[3] Das international zugängliche Europäische Genom-Phenom-Archiv (EGA) wird vom EMBL-EBI (European Molecular Biology Laboratory – European Bioinformatics Institute) in Hinxton in Großbritannien betrieben, vgl. https://ega-archive.org/.

[4] NFDI: https://www.dfg.de/foerderung/programme/nfdi/.

[5] Leipzig-Berlin-Erklärung zu NFDI-Querschnittsthemen der Infrastrukturentwicklung, veröffentlicht am 15. Juni 2020, https://zenodo.org/record/3895209#.YCJ_akBFy71

geltenden datenschutzrechtlichen Regelungen, so auch die der Datenschutz-Grundverordnung (DSGVO), wissenschaftsfreundlich ausgelegt und umgesetzt werden können, um das Forschungsprivileg der Verordnung zu verwirklichen. Neben einheitlichen, auf die Datensicherheit bezogenen Prozessen, die eine Re-Identifizierung der Betroffenen verhindern sollen, werden hierbei standardisierte Einwilligungserklärungen für das Einholen einer informierten Zustimmung der Betroffenen zu der Datenverarbeitung genannt. Außerdem sollen strukturierte Zugänge zu den Datenbeständen eingerichtet werden, die an die Sensibilität der Datentypen angepasst sind und auch eine Öffnung der Infrastrukturen für die Datennutzung im Rahmen internationaler Forschungskooperationen erlauben.

Die biomedizinische Forschung an menschlichen Genomdaten und verwandten Omics-Daten steht vor besonderen datenschutzrechtlichen Herausforderungen.[6] Menschliche Genomdaten und viele Omics-Daten fallen unter die besonderen Kategorien personenbezogener Daten, weil sie sensible Informationen über die Betroffenen vermitteln können, sowie auf der Grundlage genetischer Informationen theoretisch zur Re-Identifikation genutzt werden könnten. Neben der Identifizierung der Rechtsgrundlage für die Verarbeitung bereits vorhandener und neu erhobener Daten und der Bestimmung ihrer Standards wie etwa der Einwilligung sind weitere datenschutzrechtliche Vorgaben zu beachten, die hier lediglich beispielhaft, keineswegs vollständig hervorgehoben werden können.

In Abhängigkeit vom zur Datensicherheit eingesetzten Verfahren (z.B. Pseudonymisierung), dem Umfang der Datensätze sowie der Möglichkeit der Zusammenführung unterschiedlicher Datenbestände kann eine Re-Identifizierbarkeit der Betroffenen herbeigeführt werden. Die Verknüpfung von Datensätzen mit familiären, soziodemographischen oder audiovisuellen Informationen, die bspw. bei der Erforschung seltener Erkrankungen als Grundlage dienen, erhöht zusätzlich die Wahrscheinlichkeit einer individuellen Re-Identifikation. Die genaue Implementierung der Datenschutzprinzipien aus Art. 5 DSGVO (z.B. Datenminimierung) bedarf weiterer Präzisierung der Datenverarbeitung, insbesondere der Sicherheitsmaßnahmen, die beispielsweise anhand von Speicherungs- und Löschvorschriften geschehen kann.

Das Nutzbarmachen von Omics-Daten für die Forschung stellt GHGA vor spezifische ethische und rechtliche Herausforderungen. Die Operationalisierung der Betroffenenrechte wird daher in einem eigenen ethisch-rechtlichen Arbeitsschwerpunkt adressiert. Themen hierfür sind sowohl die sekundäre Datennutzung für Forschungszwecke und die angestrebte Weiterverwendung der Ergebnisse von Datenanalysen in der Versorgung als auch die notwendige langfristige Speicherung der Daten. Nicht zuletzt verlangt die Umsetzung der Betroffenenrechte nach einer intensiven Auseinandersetzung mit den datenschutzrechtlichen Pflichten für eine sichere und datenschutzrechtskonforme Verarbeitung. Einige dieser Pflichten ergeben sich aus der Sensibilität der zu verarbeitenden Daten, weitere Pflichten aus den infrastruktur- und konsortiumsspezifischen Daten- und Arbeitsstrukturen. In der Arbeitsstruktur muss insbesondere das Risiko der Re-Identifizierbarkeit durch Verknüpfung mit weiteren Datensammlungen oder – bei einer Einbindung der Ergebnisse in die Versorgung – mit klinischen Daten berücksichtigt werden. Auch muss bei der rechtlichen Betrachtung aufgrund unterschiedlicher beteiligter Akteure und akteursspezifischer sowie verarbeitungstypischer Sicherheitsrisiken zwischen der Archivierung inklusive der Extrahierung der Information anhand von Datenanalysen und dem Zugang zur Dateninfrastruktur sowie der Forschung mit den Daten als gesonderte Verarbeitungsschritt unterschieden werden. Das Erfordernis der Verarbeitung großer Datenmengen verlangt eine zusätzliche Auseinandersetzung mit Sicherheitsaspekten, mit den Modalitäten von Verarbeitungsvorgängen unter Beteiligung verschiedener Datenzentren sowie mit den Voraussetzungen grenzüberschreitender Datentransfers. Des Weiteren sind die Bedingungen für die Weiterverwendung der Forschungsergebnisse präzise zu bestimmen. In der Ausarbeitung der Daten- und Arbeitsstrukturen werden die FAIR-Prinzipien (*findable, accessible, interoperable, reusable*) beachtet und Interoperabilität nicht nur technisch, sondern auch juristisch umgesetzt. Dabei wird ein besonderer Fokus auf die Anschlussfähigkeit der Infrastruktur auf unionaler und internationaler Ebene gelegt.

Für das GHGA-Vorhaben ist insgesamt eine zentrale Frage, wie das Interesse der Forschung, einmal gewonnene personenbezogene und sensible Daten auf Basis der FAIR-Prinzipien[7] weiter verwendet, und die Interessen der datengenerierenden klinischen Primärversorgern mit den Datenschutzrechten der Betroffenen in der Abwägung zu einem Ausgleich gebracht werden können. Eine sichere Verarbeitung und Speicherung, klare Regelungen für den Datenzugang und für die Nutzung sowie ein begrenzter Zugriff für legitime Forschungszwecke können zu einem angemessenen Ausgleich zwischen der Forschungsfreiheit und den allgemeinen Persönlichkeitsrechten der Betroffenen maßgeblich beitragen. In dieser Abwägung kann das öffentliche Interesse an der Forschung zugunsten der Forscher berücksichtigt werden, ebenso wie das Gemeinwohl, das u.a. auf die Verbesserung der Gesundheitsversorgung durch die möglichst unmittelbare Umsetzung der Forschungsergebnisse ausgerichtet ist.

Es besteht Hoffnung, dass durch die Schaffung einer sicheren Dateninfrastruktur für diese sensiblen Daten auch das Vertrauen in diese Infrastrukturen und in die Omics-Forschung selbst wachsen wird. Die zurzeit immer noch häufig vorhandene mangelnde Akzeptanz von Datenarchiven führt häufig dazu, dass Forscher auch die rechtlich erlaubten Möglichkeiten der Datenarchivierung und des Datenaustausches sowie der Weiterverwendung der Daten für Forschungszwecke nicht ausschöpfen (können). Mit der Etablierung des GHGA und der NFDI kann dieser Tendenz entgegengewirkt werden, indem die Forschung mit Omics-Daten auf technisch und rechtlich geklärten und sicheren Grundlagen gestellt wird, sodass ihre Ergebnisse in der Tat als Basis für die weitere lebenswissenschaftliche Forschung und die Versorgung dienen können.

[6] Statt vieler vgl. Mitchell/Ordish/Johnson/Brigden/Hall, The GDPR and genomic data, Mai 2020.

[7] DFG, Die Nationale Forschungsdateninfrastruktur (NFDI), https://www.dfg.de/download/pdf/foerderung/programme/nfdi/one_pager_nfdi_de.pdf.

Dr. Fruzsina Molnár-Gábor ist Nachwuchsgruppenleiterin an der Heidelberger Akademie der Wissenschaften und am Bio-Quant-Zentrum der Universität Heidelberg. Sie war Doktorandin am Max Planck Institut für ausländisches öffentliches Recht und Völkerrecht und wurde mit einer Arbeit zur internationalen Steuerung von Genomanalysen in Heidelberg promoviert. Sie arbeitet an der Schnittstelle von Gesundheits- und Medizinrecht sowie von Datenschutzrecht und beschäftigt sich insbesondere mit der Frage, welchen Einfluss die biomedizinisch-technologische Entwicklung auf ihre Regelung hat und wie diese Entwicklung durch die Wahl verschiedener Regelungstechniken berücksichtigt und gefördert werden kann. Sie ist Mitglied der Jungen Akademie der Berlin-Brandenburgischen Akademie der Wissenschaften und der Nationalen Akademie Leopoldina, wo sie Ko-Sprecherin der Arbeitsgruppe zur Künstlichen Intelligenz ist. Die Arbeitsgruppe setzt sich zum Ziel, Fragestellungen rund um KI anhand konkreter Anwendungsfälle und korrespondierender Lösungsansätze zu diskutieren und damit den konzeptionellen Diskurs über KI zu ergänzen. Als Expertin für die rechtlichen Belange des Datenschutzes und der Datensicherheit in der medizinischen Forschung und Versorgung ist sie Mitglied zahlreicher internationaler Beratungsgremien und Forschungskonsortien wie bspw. des EU-Canada Cancer Network. Sie ist Ko-Sprecherin des GHGA-Konsortiums und leitet sein juristisches Teilprojekt. Sie ist Trägerin des Heinz Maier-Leibnitz-Preises 2020.

Dr. Jan Korbel ist Gruppenleiter und leitet die Datenwissenschaften am Europäischen Laboratorium für Molekularbiologie (EMBL) in Heidelberg. Außerdem ist er Co-Direktor der Molekularmedizin Partnerschaftseinheit der Universität Heidelberg und des EMBL. Er promovierte in Molekularbiologie an der Humboldt Universität in Berlin und arbeitete als Postdoc an der Yale University, wo er das Paired-End Mapping zur Charakterisierung struktureller Variationen durch Next-Generation Sequencing mitentwickelte. Mit seiner Expertise in Humangenetik und Computational Biology ist Jan besonders daran interessiert, die Determinanten der Bildung und Selektion genomischer DNA-Rearrangierungen (sogenannter Struktureller Varianten) zu verstehen, sowohl somatisch als auch in der Keimbahn. Als Experte leitet er federführend weltweite Konsortien in der Krebsforschung wie das Pan-Cancer Analysis of Whole Genomes Projekt. Sein Engagement gilt auch bioethischen Fragestellungen im Kontext der Ganzgenomsequenzierung bei Patienten. Jan Korbel wurde 2015 als damals eines der jüngsten Mitglieder in die Deutsche Nationale Akademie der Wissenschaften Leopoldina gewählt. Im Jahr 2018 wurde er mit dem europäischen EACR - Pezcoller Foundation Cancer Research Award ausgezeichnet. Er ist seit 2020 Teil des GHGA Direktorats (Board of Directors).