Big Data in der Genomik

Herausforderungen und Lösungen

Jedes Zeitalter hat seine technischen Durchbrüche. Die weit verbreitete Nutzung von Computern und dem Internet zu Beginn des 21. Jahrhunderts hat unsere Herangehensweise an Informationen und die Informationssuche beeinflusst [1]. Das Aufkommen sozialer Netzwerke (wie Facebook, Twitter, LinkedIn) und „Cloud"-Lösungen für die Datenspeicherung mit immer schneller werdenden Computerprozessoren hat die Art und Weise verändert, wie wir Informationen generieren [1]. Sind die Biowissenschaften auf eine Big Data-Revolution vorbereitet?

▶ **Fabrício F. Costa, Ph.D., Northwestern University's Feinberg School of Medicine, Chicago, IL, USA**

Die Biowissenschaften sind von der Erzeugung großer Datensätze stark betroffen, insbesondere durch Überladung mit so genannten „-omik"-Informationen aus biologischen Teilgebieten (Genome, Transkriptome, Epigenome und andere „-omik"-Daten von Zellen, Geweben und Organismen). Der Einsatz von DNA-Sequenzierungsmaschinen, die kleiner, aber in der Lage sind, Datenberge schneller und kostengünstiger zu generieren, hat Wissenschaft und Medizin in beispielloser Weise verändert. Die jetzige Epoche scheint das „Big Data"-Zeitalter zu werden, ein Begriff, der sich auf die explosionsartige Vermehrung der verfügbaren Informationen bezieht - ein Nebenprodukt der digitalen Revolution [2]. Während sich biomedizinische Daten in Computern und Servern rund um den Globus [1] anhäufen, werden die ersten besorgten Fragen nach der Vertraulichkeit und Sicherheit von Patientendaten gestellt.

Next-Generation Sequencing (NGS)-Plattformen, die Halbleiter [3] oder Nanotechnologie verwenden [4], haben die Generierungsgeschwindigkeit biologischer Daten in den letzten zwei Jahren exponentiell ansteigen lassen. War das erste menschliche Genom noch ein 3 Milliarden Dollar-Projekt, das sich über mehr als ein Jahrzehnt erstreckte und 2002 abgeschlossen wurde, ist man inzwischen schon fast in der Lage, ein vollständiges Genom innerhalb von wenigen Stunden für weniger als eintausend Dollar zu sequenzieren und zu analysieren. Die sinkenden Kosten haben dazu geführt, dass Informationen auf Petabyte-Ebene (10^{15} Bytes) erzeugt werden können. Computer und Internet sind zwar schneller geworden, uns fehlt jedoch die notwendige Computerinfrastruktur zur sicheren Erzeugung, Erhaltung, Übertragung und Analyse umfangreicher Informationen in den Biowissenschaften und zur Integration von „-omik"-Daten in andere Datensätze, wie zum Beispiel klinische Patientendaten (hauptsächlich aus elektronischen Patientenakten). Der vorliegende Artikel bietet einen kurzen Überblick über die Herausforderungen, mit denen wir aufgrund von Produktion, Transfer und Analyse großer Datenmengen konfrontiert sind. Außerdem wird die sich verändernde Landschaft von Privatsphäre und persönlichen Informationen im „Big Data"-Zeitalter erörtert.

Wie konnten die Datenmengen so groß werden?

Die Generierung von „Big Data" hat sich auf mehrere voneinander unabhängige Bereiche der Gesellschaft ausgewirkt, darunter Kommunikation, Medien, Medizin und wissenschaftliche Forschung [2]. In der Wissenschaft zum Beispiel wurden in weniger als 10 Jahren der Zeitaufwand und die Kosten für die Genomsequenzierung um den Faktor 1 Million gesenkt. Heutzutage können Genome für ein paar Tausend Dollar sequenziert und kartiert werden. Die persönliche Genomik ist ein Wegbereiter der prädiktiven Medizin, bei der man anhand des genetischen Profils eines Patienten die am besten geeignete medizinische Behandlung konzipieren kann.

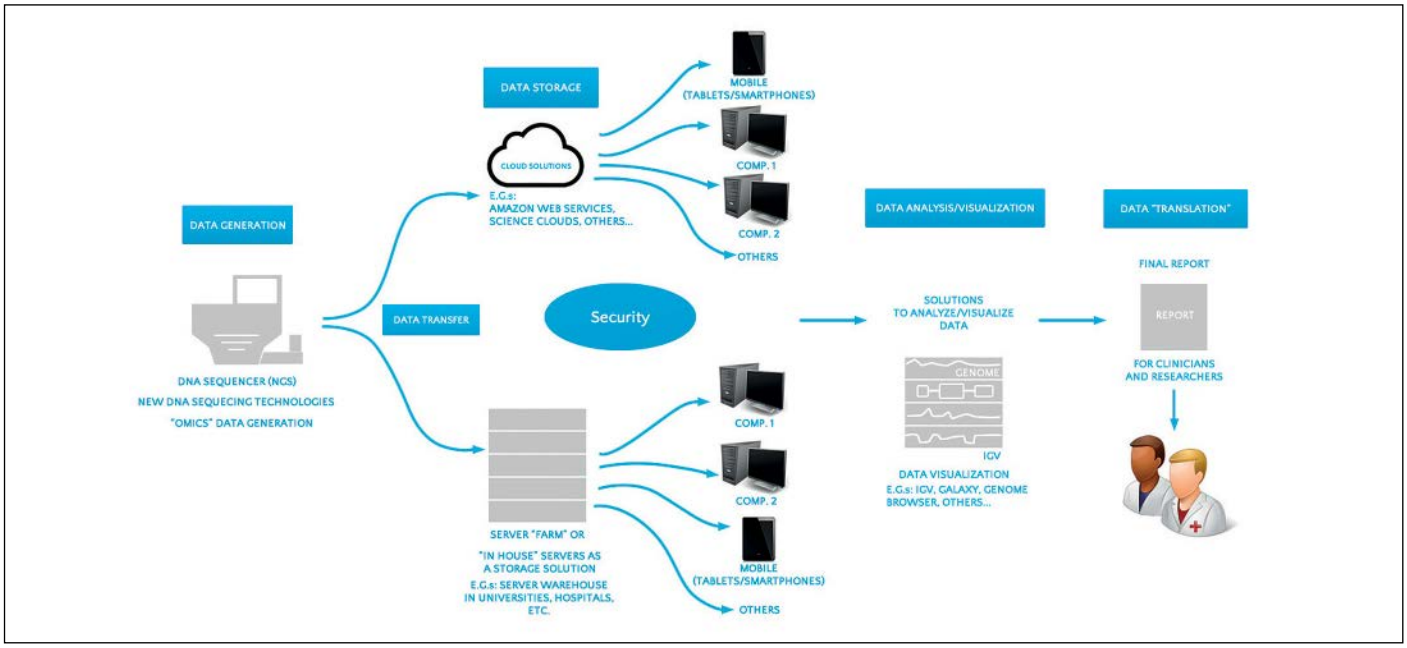

Abb. 1: Big Data in der Genomik. Schematische Darstellung Pipeline aus durch NGS generierte Daten zur Datenübertragung für Klinikärzte und Forscher. Bilddesign: Eduardo B. F. Junior. Die vollständige Beschreibung der Abbildung finden Sie unter http://goo.gl/UxSSz

Das Encode-Projekt bietet durch die Bereitstellung ausgefeilter Rahmenbedingungen für persönliche Genomik unter Mitwirkung verschiedener Forschungsgruppen eine Perspektive für die Erzeugung und Analyse großer Datenmengen [5]. Projekte wie Encode haben Berge von Daten produziert und dadurch demonstriert, wie sich Big Data zu einem festen Bestandteil der wissenschaftlichen Forschung entwickelten [5]. Tatsächlich ist die heutige Wissenschaft zunehmend „sozial", insbesondere in Bereichen wie der Genomik, in denen riesige Datenmengen erzeugt werden. Encode ist ein gutes Trainingsinstrument für Forscher in großen Wissenschaftsunternehmen, das sich verstärkt ausbreiten wird. Bei dieser Art von Projekten werden Unmengen von Daten erzeugt, gespeichert, übertragen und analysiert (für eine vollständige Übersicht siehe auch Abb. 1).

Tab. 1: Übersicht einiger Firmen, die Lösungen für die Generierung, Speicherung, Analyse und Visualisierung von „-omik"- und klinischen Daten bieten.

Firma/Institution	Lösungsmodell	Webseite
Appistry	Die Big Data-Hochleistungsplattform von Appistry kombiniert selbst-organisierende Computerspeicherung mit optimierter und verteilter Hochleistungsdatenverarbeitung zur Bereitstellung sicherer, HIPAA-konformer, akkurater On-demand-Analyse von -omik-Daten in Verbindung mit klinischen Informationen.	www.appistry.com
BGI	Die Lösung von BGI dient als solide Grundlage für großskaliges Verarbeiten von Bioinformatikdaten. Die BGI-Computerplattform ist ein integrierter Service bestehend aus vielseitig einsetzbarer Software und leistungsfähiger Hardware für die Biowissenschaften.	www.genomics.cn/en
CLC Bio	Die Firma CLC Bio hat eine Plattform, bei der sowohl Desktop- als auch Server-Software zur Erzielung besserer Leistung integriert und optimiert sind. CLC Bio verwendet auf veröffentlichten Methoden basierende selbst entwickelte Algorithmen zur erfolgreichen Beschleunigung von Datenberechnungen zur Erzielung beachtlicher Verbesserungen in der Big Data-Analyse.	www.clcbio.com
DNAnexus	DNAnexus bietet Lösungen für NGS unter Verwendung einer Cloud-basierten Datenverarbeitungs-Infrastruktur mit skalierbaren Systemen und fortgeschrittener Bioinformatik auf einer Web-basierten Plattform zur Handhabung des Datenmanagements und zur Lösung der in Verbundsystemen üblichen Analyse-Probleme.	www.dnanexus.com
Genome International Corporation	Genome International Corporation (GIC) ist ein forschungsgesteuertes Unternehmen, das innovative Bioinformatikprodukte und maßgeschneiderte Forschungslösungen für Firmen-, Regierungs- und Hochschullabors in den Biowissenschaften liefert.	www.genome.com
GNS Healthcare	GNS Healthcare ist eine Big Data-Analyse-Firma, die einen skalierbaren Ansatz für die Handhabung von Big Data-Lösungen entwickelt hat, die im gesamten Gesundheitswesen angewendet werden könnten.	www.gnshealthcare.com
Foundation Medicine	Foundation Medicine ist ein führendes Unternehmen im Bereich Molekularinformationen, das eine umfassende Krebs-Genomanalyse in den klinischen Routinebetrieb einführt. Foundation Medicine ist ein Pionier in der Entwicklung eines umfassenden Krebsdiagnosetests, der -omik-Daten, klinische Informationen und Big Data-Analytik kombiniert in der Krebsforschung einsetzt.	www.foundationmedicine.com
Knome	Knome analysiert vollständige Genomdaten unter Verwendung von Software-basierten Tests gleichzeitig, um viele Gene, Gennetzwerke und Genome zu untersuchen und zu vergleichen und um andere Formen molekularer und nicht-molekularer Daten zu integrieren. Knome bietet eine Plattform und Werkzeuge, um Forscher und Ärzte dabei zu unterstützen, Software-basierte Tests der nächsten Generation zu entwickeln und klinische Entscheidungen zu treffen.	www.knome.com
NextBio	Die Big Data-Technologie von NextBio ermöglicht den Anwendern, öffentliche und geschützte molekulare Daten und klinische Informationen von einzelnen Patienten, Bevölkerungsstudien und Modellorganismen unter Anwendung genomischer Daten in nützlicher Weise sowohl in der Forschung als auch in der klinischen Anwendung systematisch zu integrieren und zu interpretieren.	www.nextbio.com

Aufgrund des gestiegenen Bedarfs an Speicherkapazitäten für Daten und Informationen, die durch Großprojekte erzeugt wurden, sind Computerlösungen wie Cloud-basierte Datenverarbeitung entstanden. Cloud-basierte Datenverarbeitung ist das einzige Speichermodell, dass die notwendige elastische Skalierung für die DNA-Sequenzierung bieten kann, deren Entwicklungsgeschwindigkeit das Mooresche Gesetz noch übertreffen könnte. In der Tat sind Cloud-Lösungen verschiedener Firmen eingesetzt worden. Allerdings ist dies insbesondere im Hinblick auf die Sicherheit und Vertraulichkeit persönlicher medizinischer und wissenschaftlicher Daten problematisch (Abb. 1, Tab. 1). Der vielleicht größte Vorteil könnte darin bestehen, eine breite Plattform für die Entwicklung neuer Analyse- und Visualisierungsinstrumente und einen Softwareservice anbieten zu können, um diese Tools in gemeinsam genutzten Datensätzen in einem sicheren gemeinschaftlichen Arbeitsbereich zu nutzen [6]. Einige Unternehmen bieten solche Lösungen bereits an (Tab. 1). Es gäbe auch die Möglichkeit, einen App-Store insbesondere für Genomik-Tools einzurichten, auf deren Basis Hunderte von Spezialanwendungen entwickelt werden könnten [6]. Firmen wie Illumina und 23andme bieten bereits eine offene Plattform für Entwickler an, und weitere Unternehmen werden APIs (Application Programming Interfaces) in ihre Dienstleistungen integrieren. Ausschlaggebend werden jedoch Lösungen sein, die die Probleme beim Schutz privater Daten beheben.

Es werden Pipelines zur Handhabung der wachsenden Menge von Genomikdaten benötigt, um „kurze" Berichte an Forscher und Krankenhausärzte zu speichern, übertragen, analysieren, graphisch darzustellen und zu generieren (zusätzliche Informationen s. Abb. 1). Tatsächlich könnte durch die Cloud-basierte Datenverarbeitung eine völlig neue Genomikbranche entstehen, die die Medizin und die Biowissenschaften transformiert. Cloud-basierte Datenverarbeitung eröffnet der Genomikindustrie eine neue Welt von Möglichkeiten, um unsere Herangehensweise an Forschung und Medizin zu verändern. Eines der Probleme der Cloud-basierten Datenverarbeitung besteht allerdings darin, die Daten vertraulich zu behandeln.

Das nahende Zeitalter datengesteuerter Wissenschaft und Medizin

Das Verständnis dafür, wie grundlegende Systeme in lebenden Organismen funktionieren, wird die Integration vieler Schichten biologischer Informationen erforderlich machen, die von Hochleistungstechnologien erzeugt wurden. Die Komplexität der in wissenschaftlichen Projekten erzeugten Daten wird noch zunehmen, wenn wir weiterhin einzelne Zellen und Organismen isolieren und sequenzieren, während die Kosten für die Erzeugung und Analyse dieser Daten sinken, sodass mehrere hundert Millionen Proben erfasst werden können. Die Sequenzierung von DNA, RNA, dem Epigenom und anderen „-omiks" aus verschiedenen Zellen unterschiedlicher Einzelpersonen wird uns ungefähr in den nächsten 5 Jahren auf die Exabyte- (10^{18} Bytes) Datenebene katapultieren [7]. Die Integration all dieser Daten wird Hochleistungs-Computerlandschaften wie diejenigen in großen Genom-Zentren erforderlich machen [7]. Die Integration von Hardware- und Software-Infrastrukturen, die auf die Handhabung großer Datenmengen in den Biowissenschaften zugeschnitten sind, wird in den nächsten Jahren zunehmen.

Wichtig ist, dass die datengesteuerte Medizin die Entdeckung neuer Behandlungsmethoden basierend auf modellübergreifenden molekularen Messungen an Patienten und anhand der Erkenntnisse aus Tendenzen in Differentialdiagnose und -prognose sowie der Nebenwirkungen von verschreibungspflichtigen Medikamenten in klinischen Datenbanken ermöglichen wird [8]. Die Kombination von „-omik"-Daten und klinischen Patienteninformationen wird neue wissenschaftliche Erkenntnisse hervorbringen, die in den Kliniken zur Verbesserung der Patientenversorgung eingesetzt werden könnten [8]. Außerdem wird die medizinische Informatik in Form von elektronischen Patientenakten und personalisierten Therapien die Anwendung zielgerichte-

ter Behandlungen für spezifische Krankheiten ermöglichen. Es ist daher eine verlockende Vorstellung, wie sich sowohl wissenschaftliche Untersuchungen als auch die Patientenbetreuung angesichts von Big Data-Speichern verändern würden, wenn große Mengen genomischer und klinischer Daten gesammelt und vom Gesundheitspersonal gemeinschaftlich genutzt werden (Abb. 1).

Herausforderungen und Lösungen

Diese revolutionären Veränderungen bei der Erzeugung und Erfassung großer Datenmengen führen zu tiefgreifenden Herausforderungen für die Speicherung, Übertragung und Sicherheit der Informationen. Tatsächlich könnte es inzwischen kostengünstiger sein, Daten zu generieren, als sie zu speichern. Ein Beispiel dafür ist das „National Center for Biotechnology Information" (NCBI). Das NCBI ist seit 1988 mit Big Data in der Biomedizin aktiv, aber weder das NCBI noch irgendjemand im privaten Sektor verfügt über eine umfassende, kostengünstige und sichere Lösung für das Problem der Datenspeicherung (obwohl, wie in Tabelle 1 dargestellt, Firmen mit unterschiedlichen Lösungen auf den Markt kommen). Diese Kapazitäten sind für kleine Labors oder Institutionen außer Reichweite, was für die Zukunft der biomedizinischen Forschung mehrere Probleme aufwirft.

Eine weitere Herausforderung ist die Datenübertragung von einem Ort zum anderen; sie geschieht hauptsächlich durch das Versenden externer Festplatten mit der Post. Eine interessante Lösung für den Datentransfer ist die Nutzung von Biotorrents. Dieses System wird einen offenen Zugang zur gemeinsamen Nutzung wissenschaftlicher Daten ermöglichen und verwendet eine Peer-to-peer-Filesharing-Technik [9]. Torrents wurden ursprünglich konzipiert, um die Verteilung großer Datenmengen im Internet zu erleichtern, und diese Lösung könnte in der Biomedizin angewendet werden [9].

Die Sicherheit und Vertraulichkeit der Daten von Einzelpersonen ist ebenfalls problematisch. Mögliche Lösungen beinhalten den Einsatz verbesserter Sicherheitssysteme mit modernen Verschlüsselungsalgorithmen, wie sie von Banken im Finanzsektor verwendet werden, um den Datenschutz ihrer Kunden zu gewährleisten [10]. Außerdem wird eine neue Generation von Einverständniserklärungen erforderlich sein, mittels derer Studienteilnehmer oder Patienten die über sie angelegten Daten Forschern explizit offen zugänglich machen [10]. Der Einsatz von Inhouse-Hardware-Lösungen anstelle von Cloud-basierter Datenverarbeitung könnte ebenfalls zu einem verbesserten Datenschutz bei Big Data beitragen. Ein Beispiel ist das System, das Knome unter der Bezeichnung knoSYS100 einsetzt (Tab. 1). Dies sind nur einige Lösungen für die Datenschutz-Probleme im Bereich Big Data, aber es wird in naher Zukunft noch andere geben.

Schlussfolgerungen

Der Erfolg in der biomedizinischen Forschung beim Management der wachsenden Mengen von „-omik"-Daten kombiniert mit klinischen Informationen wird von unserer Fähigkeit abhängen, die hochskalierten Datensätze, die durch die neu entstehenden Technologien generiert werden, zu interpretieren. Private Unternehmen wie Microsoft, Oracle, Amazon, Google, Facebook und Twitter sind Meister im Umgang mit Datensätzen im Petabyte-Bereich. Wissenschaft und Medizin werden dieselbe Art von skalierbarer Struktur implementieren müssen, um die durch „-omik"-Techniken erzeugten Datenvolumina handhabbar zu machen. Die Biowissenschaften werden sich an die Entwicklungen in der Informatik anpassen müssen, um die Big-Data-Probleme, mit denen sie im nächsten Jahrzehnt konfrontiert sein werden, erfolgreich lösen zu können.

Danksagung

Ich bedanke mich bei Kelly Arndt und Steve Iannaccone für ihre wohl durchdachten Anregungen und das kritische Lesen dieses Artikels.

Literatur

[1] Costa F. F.: Drug Discovery Today 18, 272–281 (2013)
[2] www.nytimes.com/2012/08/12/business/how-big-data-became-so-big-unboxed.html?r=0
[3] Rothberg J. M. *et al.*: Nature 475, 348–352 (2011)
[4] Clarke J. *et al.*: Nature Nanotechnol. 4 265–270 (2009)
[5] Birney E.: Nature. 489, 49–51 (2012)
[6] www.forbes.com/sites/sap/2012/04/16/how-cloud-and-big-data-are-impacting-the-human-genome-touching-7-billion-lives/
[7] Schadt E. E. *et al.*: Nature Reviews Genetics 9, 647–657 (2010)
[8] Shah N. H. und Tenenbaum J. D.: J Am Med Inform Assoc. 19, e2–e4 (2012)
[9] Langille M. G. und Eisen J. A.: PLoS One. 5, e10071 (2010)
[10] Schadt E. E.: Mol. Syst Biol. 8, 612 (2012)

KONTAKT

Fabrício F. Costa, Ph.D.
Cancer Biology and Epigenomics Program,
Children's Hospital of Chicago Research Center
Department of Pediatrics Northwestern
University's Feinberg School of Medicine
Genomic Enterprise
Chicago, IL, USA
fcosta@luriechildrens.org
fcosta@genomicenterprise.com
fcosta@datagenno.com