Gefäßschirurgie 2019 · 24:523–530
<https://doi.org/10.1007/s00772-019-00572-9>
 Online publiziert: 17. September 2019
 © Der/die Autor(en) 2019

H. Dickten · C. Kratsch · B. Reiz

Comma Soft AG, Bonn, Deutschland

Die künstliche Intelligenz in der Einzelzellgenomik

Zwei innovative Technologien für die biomedizinische Forschung in höchster Auflösung

Die aktuellen Entwicklungen in der Digitalisierung und in den Biotechnologien werden häufig als die zwei prägenden Transformatoren für unsere Gesellschaft und unsere Zukunft genannt [1]. Während beide Entwicklungen scheinbar unabhängig verlaufen, finden sich dennoch Überschneidungen: Klassische Anwendungen der künstlichen Intelligenz (KI) halten Einzug in die medizinische Praxis, während gleichzeitig die Anforderungen der Genomik neue Algorithmen in der KI-Forschung inspirieren. Gerade in der sich derzeit etablierenden Einzelzellgenomik ist die Datenanalyse ohne KI nicht mehr denkbar.

Trend 1: Einzelzelltranskriptomik in der vaskulären Medizin

Fokus auf die kleinste Einheit des Lebens

Als kleinste Einheit des Lebens bestimmen biologische Zellen durch ihre Mechanismen und Interaktionen Struktur und Funktion aller Organismen. Störungen dieser Prozesse führen zu nichtphysiologischen Zuständen und schlussendlich zu Erkrankungen. Auch äußere Einflüsse und Umweltfaktoren werden letztlich immer auf zellulärer

Ebene verarbeitet. Dabei kann es sowohl zu aberranten spezifischen Zellzuständen als auch zur Entstehung spezifischer Zelltypen kommen [2]. Diese „Entartung von Zellen“ wird insbesondere in der Krebsforschung schon lange untersucht [3]. Aber auch vaskuläre Erkrankungen werden durch zelluläre Prozesse beeinflusst, beispielsweise bei der Entstehung von Atherosklerose, Atherothrombose, Aneurysmen, Wundheilungsstörungen oder Inflammation [4–9]. Besonders die Plastizität und Transdifferenzierung glatter Gefäßmuskelzellen (SMCs) bei der Entstehung atherosklerotischer Plaques sowie deren zelluläre Heterogenität stellt dabei ein sehr interessantes und relevantes Forschungsfeld dar [10, 11].

Neue Technologien zur Untersuchung von Zellen

Vor der Untersuchung von Genese, Progression und möglicher Behandlungen von Erkrankungen muss zunächst der physiologische Zustand eines Organismus und damit seiner Zellen und deren Interaktionen verstanden werden. In der Vergangenheit kamen dabei eine Vielzahl von Technologien als Grundlage für die systematische Beschreibung von Geweben zum Einsatz, wie zum Beispiel die Licht-/Elektronenmikroskopie, immunhistochemische Methoden wie die Durchflusszytometrie (FACS) sowie Microarrays, Fluoreszenz-in-situ-Hybridisierung (FISH) und schließlich Hochdurchsatzverfahren für die RNA- und DNA-Sequenzierung (vgl. **Tab. 1**

für zentrale Begriffe und Technologien; [2, 12]).

Ein grundlegendes Problem blieb jedoch, dass zwar unterschiedliche Zelltypen, nicht aber unterschiedliche Zellzustände differenziert werden konnten. Wesentlicher Grund dafür ist, dass sich die klassische (Bulk-)Untersuchung auf ein Gemisch von Zellen bezieht und die daraus resultierende Mittelwertmessung die wahre Zellheterogenität verschleiert. Dies änderte sich, als Expressionsprofile einzelner Zellen mittels Microarray- oder RNA-Sequenzierung zugänglich wurden (vgl. **Abb. 1**). Während zunächst nur sehr wenige Zellen untersucht werden konnten, ist es aufgrund fortschreitender Entwicklung und sinkender Kosten mittlerweile möglich, in parallelisierten Hochdurchsatzverfahren Millionen von Zellen zu untersuchen. Dabei stehen vielfältige Analysemöglichkeiten der klassischen Molekulargenetik bis hin zu Multi-OMICS-Ansätzen auch für Einzelzellanalysen zur Verfügung, was an anderer Stelle ausführlich zusammengefasst wurde [2, 3].

Einzelzelltranskriptomik als neues Forschungsgebiet

Im Folgenden beschränken wir uns auf Analysen von durch RNA-Sequenzierung gewonnenen Expressionsprofilen (Transkriptom) einzelner Zellen. Diese erlauben es nicht nur Zelltypen zu differenzieren, sondern geben auch einen Einblick in zellspezifische molekulare Prozesse. Vielfältige Studien

Die Autoren H. Dickten, C. Kratsch und B. Reiz haben zu gleichen Teilen zu der Arbeit beigetragen.

Tab. 1 Zentrale Begriffe und Technologien in der Genexpressionsanalyse

Begriff	Erläuterung
Bulksequenzierung	Bestimmung der Nukleotidabfolge von RNA einer Probe (Gewebe, Zellsuspension), die mittels Hochdurchsatzmethoden (sog. Next-Generation-Sequencing-Methoden) erfolgt. Sie gibt Aufschluss über die Funktion von Zellen
Cluster	Anonyme Klasse bzw. Gruppe von z. B. Zellen, die datengetrieben durch Gruppierung von Zellen anhand ihrer Ähnlichkeit ermittelt wird
Drop-out	Nicht detektierte Genexpression. Wird in der Genexpressionsmatrix ebenfalls mit einer Null repräsentiert
Expressionsprofil	Genexpression aller Gene einer Probe
Genexpression	Information über den Ausdruck bzw. die Erscheinung eines Gens in einer Zelle oder einem Organismus. Sie gibt Information über die Funktion einer oder mehrerer Zellen
Genexpressionsmatrix	Tabellarische Darstellung von Genexpressionen mehrerer Zellen. Da meist nicht alle Gene in allen Zellen exprimiert werden, sind diese Matrizen „dünn besetzt“ und weisen somit viele Einträge aus Nullen auf. Typische Größen von Genexpressionsmatrizen in Einzelzellstudium umfassen 100.000 Zellen × 20.000 Gene mit einer Besetzungsdichte von bis zu 5 %
mRNA	Messenger RNA oder auch Boten-RNA genannt. Abgelesener transkribierter Abschnitt der DNA, der für die Synthese von Proteinen verwendet wird
Single-Cell-Sequenzierung	Erweiterung der Bulksequenzierung zur Bestimmung der Nukleotidabfolge von RNA einer Zelle. Im Gegensatz zur Bulksequenzierung ermöglicht sie eine Auflösung auf Zellniveau
Trajektorie	Allgemein Pfad einer zeitlichen Entwicklung im Raum. Hier zweidimensionale Darstellung des Entwicklungspaths von Zellen
Transkript	Ein anhand einer gegebenen DNA synthetisiertes RNA-Molekül eines einzelnen Gens
Transkriptom	Gesamtheit aller Transkripte in einer Zelle, die exprimiert werden
Zelldifferenzierung	Entwicklungsmöglichkeit von Zellen in verschiedene Zelltypen
Zelltyp	Distinkte Einheit von Zellen, wie z. B. einer Immunzelle (B-Zellen, T-Zellen, Makrophage usw.) oder Epithelzellen
Zellzustand	Zustand einer Zelle wie aktiv, nicht aktiv oder dem Zellzyklus etc.
Zellzyklus	Abfolge von Aktivitätsphasen einer Zelle bei der Zellteilung (Mitose)

zeigen, dass diese molekulare Herangehensweise sowohl bekannte, morphologisch distinkte Zelltypen bestätigt, als auch neue (Sub-)Zelltypen identifizieren kann [13, 14]. Die Expressionsprofile einzelner Zellen können auch in ihrer natürlichen räumlichen Anordnung im Gewebe analysiert werden (sog. *spatial transcriptomics*), und es können Entwicklungswege und Übergangszustände zwischen Zellzuständen und -typen identifiziert werden (sog. *trajectories*) [15, 16]. Darüber hinaus erlaubt die hohe Sensitivität die Identifikation sehr kleiner (Sub-)Zellpopulationen. Auch *in vitro* schlecht zu kultivierende Organismen lassen sich so untersuchen, z. B. im Umgang mit Infektionskrankheiten oder beim Studium von Mikrobiomen [17]. Eine große Herausforderung liegt jedoch in der adäquaten Analyse der gewonnenen Daten, welche in aller Regel als Genexpressionsmatrizen vorliegen. Diese Matrizen sind sehr groß (>20.000 Expressionswerte pro Zelle) und nur spärlich besetzt, d. h. viele der Einträge sind „Null“, entweder da das entsprechende Gen nicht exprimiert wird, oder die Expression nicht detektiert wurde („Dropout“). Ein Hilfsmittel, um die

se Herausforderungen anzugehen, ist die künstliche Intelligenz, auf die wir im weiteren Verlauf des Artikels näher eingehen werden (s. Trend 2).

Medizinische Forschung in der Einzelzelltranskriptomik

Ein Treiber der Einzelzelltranskriptomik ist die Krebsforschung. Hier ermöglicht die Einzelzellsequenzierung u. a. die Katalogisierung von Tumor-Teilpopulationen, eine Detail-Charakterisierung von Tumorzellen und deren zeitliche und phylogenetische Entwicklung, die Detektion seltener Zellen und das Studium der Interaktion eines Tumors mit seiner Umgebung. Ähnlich relevante Anwendungsmöglichkeiten ergeben sich für das Studium des zentralen Nervensystems, des Immunsystems oder für die Untersuchung der frühen Embryonalentwicklung bei Säugetieren [7, 18, 19].

Besonders hervorzuheben ist hier der „Human Cell Atlas“ [2], ein aktuelles Forschungsprojekt zur Katalogisierung aller Zelltypen des gesunden menschlichen Körpers. Verwendet man diese Zelltypen als Referenz für anschließende Studien, u. a. im Krankheitskontext, können die

Erkenntnisse aus dem Bereich der Einzelzellanalyse auch die Diagnostik verbessern, beispielsweise durch neu identifizierte Biomarker.

» Ein Treiber der Einzelzelltranskriptomik ist die Krebsforschung

Im Bereich vaskulärer Erkrankungen liegt der Fokus der Transkriptomanalyse von einzelnen Zellen aktuell im Bereich der Inflammation und Atherosklerose. Das Potenzial zeigte sich bereits in ersten Studien zur Heterogenität von Makrophagen der Aorta, der Identifikation distinkter Endothelzellpopulationen, deren Hierarchie sowie Übergangszuständen in der Aorta, der zellulären Zusammensetzung von Blutgefäßen im Gehirn und der Lunge und dem Immunzellrepertoire im Atherosklerose-Mausmodell [5, 20–22]. Besonders relevante und neue Erkenntnisse durch Einzelzellanalysen liefern hochaktuelle Studien zur Entstehung, Zusammensetzung und der zellulären Plastizität in atherosklerotischen Plaques [23]. Von herausragender Bedeutung ist dabei

Gefäßchirurgie 2019 · 24:523–530 <https://doi.org/10.1007/s00772-019-00572-9>
© Der/die Autor(en) 2019

H. Dickten · C. Kratsch · B. Reiz

Die künstliche Intelligenz in der Einzelzellgenomik. Zwei innovative Technologien für die biomedizinische Forschung in höchster Auflösung

Zusammenfassung

Die individuelle Zelle stellt die fundamentale Einheit des Lebens dar. Ihre Funktionsweise ist seit Jahrhunderten Gegenstand der biomedizinischen Forschung. Seit Kurzem erlaubt die sogenannte Einzelzellsequenzierung revolutionäre neue Einblicke in die Komposition von Geweben, die Interaktion zwischen Zellen und den Ablauf dynamischer Prozesse. Die damit verbundenen Daten, z. B. aus der Transkriptomik, stellen Analysten jedoch vor neue Herausforderungen. Die Daten sind typischerweise sehr groß, verrauscht und stark mit anderen Informationen vernetzt, sodass herkömmliche Verfahren ausscheiden. Neue,

speziell adaptierte Algorithmen aus dem Feld der künstlichen Intelligenz können hier Abhilfe schaffen. In Verbindung stellen Einzelzellsequenzierung und künstliche Intelligenz mächtige neue Werkzeuge für die biomedizinische Forschung dar und erlauben Einblicke in höchster Auflösung. Wir geben einen kurzen Überblick über beide Technologien und liefern Beispiele für ihre Anwendung im medizinischen Bereich. Wir demonstrieren anschließend beispielhaft, wie Methoden der künstlichen Intelligenz erfolgreich für die Analyse von Einzelzelltranskriptomdaten eingesetzt

werden können. Da der korrekte und gewinnbringende Einsatz dieser Methoden weiterhin detailliertes Hintergrundwissen verlangt, ist ein intensiver Austausch zwischen den Disziplinen unumgänglich. Wir gehen deshalb abschließend auf den Einsatz neuer Plattformen zur Kollaboration und zum Wissensaustausch ein.

Schlüsselwörter

Genomik · Zelle · Künstliche Intelligenz · Deep Learning · Machine Learning

Artificial intelligence in single cell genomics. Two innovative technologies for biomedical research in highest resolution

Abstract

The individual cell represents the fundamental unit of life. Its manner of functioning has been the focus of biomedical research for centuries. In recent years, advances in high throughput so-called single cell sequencing techniques have made it possible to study individual cells and their genetic profile. This enables revolutionary new insights into tissue composition, cell–cell interactions and dynamic processes in health and disease. The resulting profile data, e.g. from single cell transcriptomics, however, provide analysts with new challenges: data sets are typically very large, noisy and highly interconnected with other annotation data, making them

unsuitable for established procedures. The setting calls for the application of novel algorithms originating from the field of artificial intelligence, which are adapted to deal with this type of challenge. Together, single cell sequencing and artificial intelligence can be considered powerful tools for biomedical research, enabling insights at the highest resolution. This article provides a short overview of the recent developments in both technologies and gives examples for their impact on medical applications. Subsequently, it is demonstrated how methods from artificial intelligence can be successfully applied for the analysis of

single cell transcriptomics data. Since the successful application of such methods still requires a detailed understanding of their requirements, an even stronger interaction between specialists of both disciplines may become necessary in the future. We therefore conclude this article with a comment on the use of new platforms for collaboration and knowledge exchange.

Keywords

Genomics · Cells · Artificial intelligence · Deep learning · Machine learning

die phänotypische Modulation (Phänotypswitch) glatter Gefäßmuskelzellen (SMCs) und deren Fähigkeit auf verschiedenste Weise transdifferenzieren zu können, um beispielsweise gewisse Subtypen von Schaumzellen zu bilden [11]. Zudem konnte eine kürzlich erschienene Einzelzellstudie erstmals die phänotypische Modulation von SMCs mit einer durch genomweite Assoziationsstudien (GWAS) identifizierten Variante für die Koronare Herzkrankheit (KHK) in Zusammenhang gebracht werden [24, 25]. Die gefundene Variante im *TCF21*-Gen ist dabei protektiv und führt zur Differenzierung der SMCs

in einen fibroblastenähnlichen Zelltyp („Fibromyocyten“ genannt) anstelle eines klassischen Makrophagenphänotyps. Weitere Forschung auf diesem Feld wird in naher Zukunft erstmals die Erstellung eines umfassenden Zellatlases der Atherosklerose erlauben [26].

Trend 2: Künstliche Intelligenz als Wegbereiter für Medizin und Forschung

KI im Alltag und speziell in der Medizin

Methoden der künstlichen Intelligenz (KI) halten mit rasanter Geschwindigkeit Einzug in unseren Alltag: Das Spektrum von Anwendungen reicht dabei von Produktempfehlungen, Preisgestaltung, Chatbots und Spam-Erkennung über Spracherkennung, Navigation und autonomes Fahren bis hin zu Mustererkennung in Sensordaten, Bildern und

Abb. 1 ▲ Durch die höhere Auflösung ergeben sich umfangreiche zusätzliche Erkenntnisse. Während bei herkömmlichen Bulk-Verfahren, unter anderem beim Vergleich zweier Proben (z. B. Test- vs. Kontrollgruppe), nur die zwei Mittelwerte aller Zellen der jeweiligen Proben verglichen werden können, liefern Einzelzell-Transkriptomik-Studien eine Vielfalt an weiteren Vergleichsmöglichkeiten

Filmen. In den verschiedensten Bereichen von Wirtschaft und Forschung wird künstliche Intelligenz vermehrt zur Automatisierung von repetitiven Aufgaben eingesetzt und ist somit ein wesentlicher Treiber der digitalen Transformation [27].

Getrieben wird dieser Prozess durch das Entstehen großer annotierter Datenräume (Big Data), innovativer Analysemethoden (KI, maschinelles Lernen, „deep learning“), und neuer technischer Möglichkeiten in der Berechnung komplexer Algorithmen und Analyseaufgaben (GPU-Beschleunigung, Cloud-Computing, Big-Data-Technologien). Die hier genannten Begriffe werden aus diesem Grund jedoch häufig fälschlicherweise synonym verwendet, obwohl eine klare Abgrenzung möglich und nötig ist (■ Tab. 2).

Auch die moderne medizinische Versorgung bleibt nicht unbeeinflusst von der Kombination aus digitaler Transformation und Fortschritten in den Biotechnologien: Das Spektrum der Durchbrüche, erzielt durch Einsatz künstlicher Intelligenz, reicht dabei derzeit von (i) beschleunigter oder verbesserter Verarbeitung oder Interpretation von bildgebenden Verfahren über (ii) Verbesserung von Prozessen und Auswertungsverfahren durch Automatisierung und Fehlerreduktion bis hin zu (iii) der verstärkten Einbindung und Ermächtigung des Patienten durch digitale Systeme.

KI erlaubt erstmals die automatisierte, objektive und reproduzierbare Hautkrebsfrüherkennung sowie Identifikation von Läsionen und Tumoren. Außerdem können Operationen in zuvor undenkbarer Präzision durch Verarbeitung und Einblendung von Echt-

zeitinformationen (*augmented reality*) oder auch Roboterunterstützung realisiert werden. Weiterhin erlaubt die KI Patienten verstärkt durch Wearables und andere digitale Systeme den eigenen Gesundheitszustand zu verfolgen, kommunizierbar zu machen und zusammen mit den entsprechenden Experten Maßnahmen zu treffen. Schlussendlich besteht die Hoffnung, mittels KI die Zeit bis zur Diagnose zu verkürzen, die Fehlerrate zu verringern und schließlich die Versorgung zu verbessern. Eine exzellente und ausführliche Zusammenstellung und Auswertung weiterer Beispiele wurde erst vor Kurzem an anderer Stelle veröffentlicht [28].

Vorzüge von KI-Systemen gegenüber menschlichen Experten

Auch wenn die Anwendungsgebiete und Aufgabenstellungen für den Einsatz von KI-Systemen sehr vielfältig sind, so lassen sich häufig auf einem abstrakten Level klare gemeinsame Vorteile der digitalen Automatisierung erkennen. Erfolgreiche Einsätze von KI in der Medizin verbessern üblicherweise die Qualität oder den Zugang zu wichtigen objektiv messbaren Kenngrößen (z. B. bei der Detektion, Lagebestimmung oder Kategorisierung) gegenüber herkömmlichen Ansätzen durch menschliche Experten. Dies lässt sich zum einen darauf zurückführen, dass KI-Systeme für eine solche Aufgabenstellung hoch spezialisiert sind und deterministische, reproduzierbare Ergebnisse liefern; zum anderen werden diese Expertensysteme anhand tausender bis Millionen von Beispielen trainiert und verifiziert, während ein durchschnittlicher menschlicher Experte naturgemäß eine deutlich geringere Menge an Fällen in seinem Berufsleben diagnostiziert.

Gleichzeitig weist einmal maschinell inkorporiertes Expertenwissen einen weiteren entscheidenden Vorteil auf: Die trainierten mathematischen Modelle lassen sich leicht vervielfältigen, konservieren und transportieren und können somit flächendeckend bis an Orte mit erschwerten Bedingungen und

Tab. 2 Zentrale Begriffe und Technologien im KI-Umfeld

Begriff	Erläuterung
Big Data	Sammelbegriff für Algorithmen (z. B. ML-Verfahren) und Softwarelösungen (z. B. Datenbanken), welche mit extrem großen, heterogenen und ggf. verteilten Datenmengen umgehen können
Cloud-Computing	Sammelbegriff für Software- und Hardware-Lösungen sowie Geschäftsmodelle, welche digitale Prozesse in standardisierter Weise in externe Rechenzentren auslagern
Deep learning	Besondere Art von maschinellen Lernverfahren, welche auf Basis von neuronalen Netzen arbeitet. Häufig besonders erfolgreich, wenn große Datenmengen zum Lernen verfügbar sind
Feature	Messbare Größe oder Eigenschaft observierter Systeme anhand derer Modelle trainiert werden. Dies kann eine Spalte einer gemessenen Datenmatrix (z. B. Höhe, Breite, Gewicht, Farbe) oder eine zusätzliche daraus abgeleitete Größe sein (z. B. Dichte)
Künstliche Intelligenz (KI)	Im weitesten Sinne: Automatisierung von intelligentem Verhalten. Im Alltagsgebrauch bezeichnet der Begriff die digitale Automatisierung von Prozessen, insbesondere von repetitiven Aufgaben des menschlichen Handelns und Entscheidens
Maschinelles Lernen (ML)	Maschinelles Lernen bezeichnet eine Reihe von mathematischen Verfahren, welche Muster, Regeln oder Entscheidungen aus Daten extrahieren und reproduzieren. Essenzielles Werkzeug der KI
Modell	Das trainierte Regelwerk: Ergebnis des Algorithmus gepaart mit Trainingsdaten. Dies kann sowohl ein künstliches neuronales Netzwerk sein, dessen Ausmaß und Kopplungsgewichte durch das Training gesetzt sind als auch ein Entscheidungsbaum. Durch Anwendung des Modells auf neue Daten kann eine zu den Trainingsdaten analoge Ausgabe bzw. Transformation erzeugt werden
(Un-)Supervised learning	Technik zur Erzeugung mathematischer Modelle mithilfe von Algorithmen und Trainingsdaten. Beim „supervised learning“ wird zu den Daten das gewollte Verhalten/die Ausgabe vorgegeben; beim „unsupervised learning“ erfolgt die Erzeugung rein datengetrieben

medizinischer Unterversorgung zur Verfügung gestellt werden.

Herausforderungen beim Einsatz von KI

Beim Einsatz von KI-Methoden wird häufig maximale *Transparenz* über die Entscheidungsfindung gefordert. Dies wird von KI-Experten leider gelegentlich missverstanden als eine Forderung nach völliger Nachvollziehbarkeit und Dokumentation der (oft höchst komplexen) mathematischen und algorithmischen Abläufe. Das Vertrauen in KI beim Anwender steht und fällt jedoch vor allem mit einem Verständnis über die Chancen und Möglichkeiten, aber auch Grenzen der jeweiligen Algorithmen und Systeme sowie der Darreichungsform der Ergebnisse von solchen Entscheidungs- und Assistenzsystemen. So können unabhängig von der Aufgabenstellung trainierte KI-Systeme nicht „zwischen den Zeilen lesen“ und somit nicht intuitiv entscheiden, wie es ein menschlicher Experte kann. Sie verarbeiten lediglich exakt die Informationen, auf denen zuvor trainiert wurde, wohingegen ein erfahrener Chirurg beispielsweise einen Tumor erkennen kann, auch wenn er eine gewisse Tumorart an dieser Stelle im Körper zuvor noch nie gesehen hat.

Einsatz von KI in der medizinischen Forschung

Neben den oben aufgezeigten Einsatzgebieten in Diagnostik und Automatisierung findet KI in jüngster Zeit auch vermehrt Anwendung in medizinischer Grundlagenforschung zur Aufklärung von Pathogenese und der Identifikation neuer geeigneter Biomarker. So können beispielsweise in der Neurologie durch Analyse von Bewegung, Sprache und Schriftbild kleinste Veränderungen mittels des Einsatzes von KI aufgedeckt werden und durch Identifikation korrespondierender *Features* (■ **Tab. 2**) in Biomarker überführt werden. Ein ähnliches Vorgehen findet in der Einzelzellanalytik von Transkriptomdaten Anwendung. Hier werden mittels KI-gestützter Methoden die Funktion und Heterogenität von Zellen untersucht, um damit Informationen über Pathogenese und Wirkmechanismen komplexer Krankheitsbilder wie Inflammation und Atherosklerose liefern zu können [2, 5, 29].

Obleich die Anwendungsgebiete von KI in der medizinischen Forschung sehr vielfältig sind, lässt sich eine typische Vorgehensweise mittels *Kernthemen der KI* darstellen. So werden die häufig erheblichen Mengen an Bilddaten oder Zeitreihen von Sensordaten zunächst

auf die wesentlichen Informationen reduziert (*Dimensionsreduktion*), Ähnlichkeiten hervorgehoben (*Clustering* und *Mustererkennung*) und Auffälligkeiten sowie Abweichungen zu erwartetem Verhalten (*Modellbildung: Training und Vorhersage*) für die Experten kenntlich und zugänglich gemacht.

Methoden der KI im Bereich der Einzelzellanalytik

Die Kombination der Trends

Aktuelle Studien in der Einzelzelltranskriptomik befassen sich mit der Beschreibung der Zellvielfalt in einer oder mehreren Stichproben, der vergleichenden Analyse zwischen zwei oder mehr Konditionen (z. B. Test- vs. Kontrollgruppe), der Untersuchung des Einflusses der Umgebung der Zelle im Gewebe, aber auch der Entwicklung von Zellen im Rahmen eines zeitlichen oder dynamischen Prozesses wie der Zelldifferenzierung oder des Zellzyklus. Analyse-Pipelines in der Einzelzellforschung weisen deshalb einen typischen Ablauf auf (vgl. ■ **Abb. 2**).

Die Kombination aus den typischerweise sehr großen Datenmengen (*volume*), den Schwierigkeiten durch technische Artefakte (*veracity*) und der Vielfalt an Daten durch *Multi-omics*-Perspekti-

Abb. 2 ▲ Typischer Ablauf der Analyse von Einzelzelltranskriptomdaten. Eine Auswahl der hier genannten Schritte ist der Kern jeder Einzelzell-Transkriptomik-Studie

ven (*variety*) macht Einzelzelltranskriptomik dabei zu einem typischen Big-Data-Problem. Es ist daher wenig überraschend, dass Einzelzellanalysen ein typischer Anwendungsfall für die *Kernthemen von KI* sind. Im Folgenden sollen zur Demonstration exemplarisch vier wesentliche Schritte einer Einzelzellanalyse sowie die dabei verwendeten KI-Methoden erläutert werden.

Beispiel 1: Vereinfachung der Datenräume (Dimensionsreduktion)

Hochdimensionale Count-Matrizen, welche Daten zu tausenden Genen in zehntausenden biologischen Zellen enthalten, sind inhaltlich oft redundant und für Menschen schwer oder gänzlich unlesbar. KI-Verfahren zur *Dimensionsreduktion* vereinfachen die Daten selbstständig bei möglichst geringem Informationsverlust, indem sie die Koordinaten in neuen Formaten darstellen, sodass eine grafische Darstellung in 2-D oder 3-D (anstatt der tausenden Dimensionen) möglich wird (vgl. ■ **Abb. 2**, Mitte oben). Für Einzelzelldaten kommen dabei KI-Algorithmen mit Namen wie *t-SNE* oder *UMAP* zum Einsatz, sowie sog. *Autoencoder* aus dem Bereich

des „*deep learning*“ (vgl. Überblick z. B. in [30]).

Beispiel 2: Gruppierung (Clustering)

Ein naher verwandter Arbeitsschritt ist die Gruppierung von Zellen zu anonymen Klassen gemäß systematischer Ähnlichkeiten. KI-Verfahren zum sogenannten *Clustering* gruppieren selbstständig biologische Zellen auf Basis eines definierten Verständnisses von Ähnlichkeit. So könnte z. B. die Klasse aller kurz vor dem Zelltod befindlichen Zellen oder die Klasse aller Zellen mit hoher Aktivität in bestimmten Genen gänzlich automatisiert entstehen. Sehr häufig finden sich dutzende anonyme Klassen in einer Stichprobe. Anschließend können statistische Eigenschaften der Klassen (z. B. Größe, Variabilität, klassentypische Markergene und klassenspezifische biomolekulare Prozesse) von menschlichen Experten begutachtet werden, um die Klassen zu benennen (vgl. ■ **Abb. 2**, rechts oben). Häufig lassen sich so bereits längst bekannte Zelltypen identifizieren, aber auch neue, unbekannten Typen, Subtypen oder Zwischenstufen finden. Typische Algorithmen für das *Clustering* sind generische Verfahren, die auch in anderen Wissenschaften und Industriebran-

chen zum Einsatz kommen, wie *hDBSCAN* oder *k-Means*, oder einzelzellspezifische Methoden wie der Louvain-Algorithmus [31].

Beispiel 3: Zeitliche Entwicklung (Mustererkennung)

Neben dem Zelltyp als Identität einer biologischen Zelle stellt ihre Position in einer zeitlichen Entwicklung oder in einem dynamischen Prozess eine weitere Facette dar. Beispielsweise mag man sich fragen, ob sich die Zellen einer Stichprobe eher am Anfang, in der Mitte oder am Ende des Zellteilungsprozesses befinden. Oder man fragt sich grundsätzlich, nach welchen Regeln sich Stammzellen hin zu spezifischen Zelltypen entwickeln. Frühe Studien lieferten bereits Hinweise auf eine komplexe Landschaft an Zellzuständen [2]. Mittlerweile existiert eine Vielzahl an spezialisierten KI-Verfahren, welche sogenannte *Trajektorien* (d. h. zeitliche Entwicklungspfade) automatisiert aus Daten bestimmen (Überblick in [32]), und damit Erkenntnisse über die Dauer von Veränderungen oder die logische Anordnung von Vorgängerzellen oder -zuständen liefern können (vgl. ■ **Abb. 2**, links unten).

Beispiel 4: Zelltypisierung (Modellbildung)

Die Benennung der Zellklassen durch Zuweisung zu bekannten Zelltypen erfolgt im Allgemeinen manuell durch Experten (vgl. [Abb. 2](#), rechts oben). Dieser Ansatz weist jedoch eine hohe Subjektivität auf, ist mit hohem Zeitaufwand verbunden und stößt bei sehr nah verwandten Zelltypen an seine Grenzen. Durch die derzeit entstehende Vielzahl an neuen Referenzdatensätzen [2] wird jedoch der Einsatz von automatisierten KI-Verfahren zunehmend machbar. Solche Verfahren können das Expertenwissen zu den bekannten Eigenschaften etablierter Zelltypen erlernen, indem sie automatisiert Wissen aus den Referenzdatensätzen extrahieren (= „Modelle bilden“). Das gelernte Wissen kann dann auf neue Datensätze übertragen werden. So kann in einem neuen Datensatz jeder einzelnen Zelle anhand ihres individuellen Profils der wahrscheinlichste Zelltyp präzise zugewiesen werden, ohne dass menschliche

Experten erneut bemüht werden müssen [33].

Kollaboration zwischen den Wissenschaften

Wie oben beschrieben, bedeutet die Durchführung einer Einzelzellstudie die Analyse von mehreren tausend Zellen, also großen und vernetzten Datenmen-gen, verbunden mit der Anwendung einer Vielzahl von frei parametrisierbaren Algorithmen. Dafür sind eine umfassende Expertise in der Programmierung und Datenmanipulation sowie große Rechnerressourcen erforderlich. Da sich darüber hinaus das junge Feld der Einzelzellanalytik noch sehr schnell weiterentwickelt und regelmäßig neue Methoden entstehen, sind *Best Practices* und einfache Software selten. Neben dem intensiven Austausch zwischen klinischem Bereich und dem *wet lab* wird deshalb eine intensive Zusammenarbeit mit Statistikern, Algorithmikern und

Ingenieuren in Zukunft zunehmende Bedeutung zukommen.

Ergänzend dazu entstehen derzeit interaktive, webbasierte Analyse-Tools in der Cloud wie *FASTGenomics* und *Gra-natum* [34, 35], welche neben einem grafischen Interface auch die notwendigen Rechenressourcen komfortabel für das Studium eigener Datenmengen zur Verfügung stellen. Die Autoren haben vor kurzem eine Studie zum Tool *FASTGenomics* [34] vorgelegt, welches als KI-Plattform zum einen etablierte Workflows zur KI-getriebenen Analyse von Einzelzelltranskriptomdaten bietet, als „collaboration platform“ zum anderen aber auch den Austausch zwischen den Experten-richtungen explizit unterstützt.

Fazit für die Praxis

- **Einzelzellgenomik ist ein schnell voranschreitender Trend zur Analyse biomolekularer Vorgänge in Zellen in höchster Auflösung.**

Hier steht eine Anzeige.

- **Künstliche Intelligenz** hilft nicht nur in der Genomforschung, sondern in vielen Anwendungsbereichen der Biomedizin und Gesundheitsversorgung bei der Aufbereitung von digitalen Daten.
- In Kombination liefern KI-basierte Algorithmen in der Einzelzelltranskriptomik Mehrwerte z. B. in der Vereinfachung von Daten oder der datengetriebenen Zelltyp- oder Zellzustandsbestimmung.
- Beide Technologien, Einzelzellgenomik und künstliche Intelligenz, benötigen tiefes Expertenwissen und können nur im Austausch und intensiven Zusammenspiel beider Disziplinen erfolgreich sein.
- Zum aktuellen Zeitpunkt erfordern die genannten Methoden häufig umfassende Programmierkenntnisse.
- KI-Plattformen bieten eine gute Voraussetzung für die gemeinsame Arbeit.

Korrespondenzadresse

Dr. C. Kratsch
Comma Soft AG
Pützchens Chaussee
202–204a, 53229 Bonn,
Deutschland
christina.kratsch@
comma-soft.com

Einhaltung ethischer Richtlinien

Interessenkonflikt. H. Dickten, C. Kratsch und B. Reiz sind Mitarbeiter der Firma Comma Soft AG, die das Projekt *FASTGenomics* in Kooperation mit dem LIMES Institut der Universität Bonn durchführt. *FASTGenomics* – Erstellung einer Big Data Analytics Plattform für die Einzelzelltranskriptomik – wird gefördert durch das Bundesministerium für Wirtschaft und Energie aufgrund eines Beschlusses des Deutschen Bundestages.

Für diesen Beitrag wurden von den Autoren keine Studien an Menschen oder Tieren durchgeführt. Für die aufgeführten Studien gelten die jeweils dort angegebenen ethischen Richtlinien.

Open Access. Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (<http://creativecommons.org/licenses/by/4.0/deed.de>) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Literatur

- Harari YN (2018) Homo deus – Eine Geschichte von Morgen. Aus dem Englischen von Andreas Wirthensohn. C.H.Beck, München
- Regev A, Teichmann SA, Lander ES et al (2017) The human cell atlas. *Elife* 6:e27041
- Baslan T, Hicks J (2017) Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat Rev Cancer* 17:557
- Freedman JE (2011) A platelet transcriptome revolution. *Blood* 118:3760–3761
- Winkels H, Ehinger E, Goshesh Y et al (2018) Atherosclerosis in the single-cell era. *Curr Opin Lipidol* 29:1
- Diamond SL (2016) Systems analysis of thrombus formation. *Circ Res* 118:1348–1362
- Lintermans LL, Stegeman CA, Heeringa P, Abdulahad WH (2014) T cells in vascular inflammatory diseases. *Front Immunol* 5:504
- Michel J-B, Li Z, Lacolley P (2012) Smooth muscle cells and vascular diseases. *Cardiovasc Res* 95:135–137
- Quintana RA, Taylor WR (2019) Cellular mechanisms of aortic aneurysm formation. *Circ Res* 124:607–618
- Gomez D, Owens GK (2012) Smooth muscle cell phenotypic switching in atherosclerosis. *Cardiovasc Res* 95:156–164
- Owsiany KM, Alencar GF, Owens GK (2019) Revealing the origins of foam cells in atherosclerotic lesions. *Arterioscler Thromb Vasc Biol* 39:836–838
- Lander ES (1996) The new genomics: global views of biology. *Science* 274:536–539
- Shekhar K, Lapan SW, Whitney IE et al (2016) Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 166:1308–1323.e30
- Zeisel A, Muñoz-Manchado AB, Codeluppi S et al (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347:1138–1142
- Burgess DJ (2019) Spatial transcriptomics coming of age. *Nat Rev Genet* 20:317–317
- Saelens W, Cannoodt R, Todorov H, Saeys Y (2019) A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 37:547–554
- Illumina An Overview of Recent Single-Cell Research Publications Featuring Illumina® Technology 3. https://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/single-cell-sequencing-research-review.pdf. Zugegriffen: 24.6.2019
- Griffiths JA, Scialdone A, Marioni JC (2018) Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol Syst Biol* 14:e8046
- Strzelecka PM, Ranzoni AM, Cvejic A (2018) Dissecting human disease with single-cell omics: application in model systems and in the clinic. *Dis Model Mech* 11:dmm36525
- Cochain C, Vafadarnejad E, Arampatzis P et al (2018) Single-cell RNA-seq reveals the transcriptional landscape and heterogeneity of aortic macrophages in murine atherosclerosis. *Circ Res* 122:1661–1674
- Lukowski SW, Patel J, Andersen SB et al (2019) Single-cell transcriptional profiling of aortic endothelium identifies a hierarchy from endovascular progenitors to differentiated cells. *Cell Rep* 27:2748–2758.e3
- He L, Vanlandewijck M, Mäe MA et al (2018) Single-cell RNA sequencing of mouse brain and lung vascular and vessel-associated cell types. *Sci Data* 5:180160
- Dobnikar L, Taylor AL, Chappell J et al (2018) Disease-relevant transcriptional signatures identified in individual smooth muscle cells from healthy mouse vessels. *Nat Commun* 9:4567
- Wirka RC, Wagh D, Paik DT et al (2019) Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. *Nat Med* 25(8):1280–1289. <https://doi.org/10.1038/s41591-019-0512-5>
- Pan H, Reilly MP (2019) A protective smooth muscle cell transition in atherosclerosis. *Nat Med* 25(8):1194–1195. <https://doi.org/10.1038/s41591-019-0541-0>
- Zhang H, Zhang NR, Li M, Reilly MP (2018) First giant steps toward a cell atlas of atherosclerosis. *Circ Res* 122:1632–1634
- Ligeza A (1995) Artificial intelligence: a modern approach. *Neurocomputing* 9:215–218
- Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 25:44–56
- Schultze JL (2019) Myocardial infarction cell by cell. *Nat Immunol* 20:7–9
- McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. *J Open Source Softw* 3:861
- Blondel VD, Guillaume J-LL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008:P10008
- Wolf FA, Angerer P, Theis FJ (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19:15
- Aran D, Looney AP, Liu L et al (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 20:163–172
- Scholz CJ, Biernat P, Becker M et al (2018) FASTGenomics: an analytical ecosystem for single-cell RNA sequencing data. *bioRxiv* 272476
- Zhu X, Wolfgruber TK, Tasato A et al (2017) Granatum: a graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med* 9:108