GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Abschlussbericht

Zuwendungsempfänger:	Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt GmbH Institute of Computational Biology (ICB)
Projektleiter:	Dr. Matthias Heinig
Projekttitel:	VALE: **V**ariant calling and effect prediction by **A**rtificial intelligence for **LE**ukemia diagnosis and subtype identification Entdeckung und Vorhersage der Wirkung von genetischen **V**arianten durch **A**rtifizielle Intelligenz für **LE**ukämie Diagnose und Subtyp-Identifizierung
Förderkennzeichen:	**031L0203A**
Laufzeit des Projektes:	01.01.2020 - 30.11.2023

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

I. Kurze Darstellung

1. Aufgabenstellung,

Leukämie ist eine wichtige Gesundheitsbelastung und eine sehr heterogene Erkrankung. Obwohl die genaue molekulare Charakterisierung des Krebs-Subtyps für die Behandlung des Patienten von entscheidender Bedeutung ist, ist ein Großteil der genetischen und molekularen Heterogenität des Patienten nicht bekannt. Um auf diese Bedürfnisse einzugehen, haben wir ein neues Forschungsnetzwerk aufgebaut, VALE (**V**ariant calling and effect prediction by **A**rtificial intelligence for **LE**ukemia diagnosis and subtype identification), das drei Forschungsgruppen mit komplementären Fachwissen und Ressourcen zusammenbringt, um diese Herausforderung zu bewältigen: i) Julien Gagneur hat den Nachweis der aberranten Expression zur Diagnose von seltenen Krankheiten vorangetrieben; ii) Matthias Heinig entwickelt Berechnungsmethoden, um nicht-kodierende genetische Variation zu interpretieren; iii) Stephan Hutter, von MLL, betreut eine einzigartige Genomressource, die mehr als 4.000 vollständigen Genomen (WGS, Whole Genome Sequencing) und entsprechenden Transkriptomen (RNA-Seq) von Leukämiepatienten sowie weiteren Informationen zu Behandlungsergebnissen und Überleben umfasst.

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde,

Das Projekt wurde im Rahmen des “Computational Life Science” Programmes durchgeführt. Dieses Programm fördert die Entwicklung innovativer Methoden und Software-Werkzeuge aus Bioinformatik, Modellierung und Simulation für den Einsatz in den Lebenswissenschaften. Spezifisch wurde in der zweiten Auswahlrunde die Entwicklung von Deep Learning Methoden in den Lebenswissenschaften gefördert.

3. Planung und Ablauf des Vorhabens

Um eine möglichst genaue Diagnose und zielgerichtete Therapie zu ermöglichen, ist das Münchner Leukämielabor (MLL) führend in der routinemäßigen klinischen Diagnostik. Darüber hinaus verfügt das MLL über eine einzigartige Genomressource mit mehr als 5.000 vollständigen Genomen (WGS, Whole Genome Sequencing) und passenden Transkriptomen (RNA-Seq) von Leukämie-Patienten sowie Follow-up-Informationen zu Behandlungsergebnissen und Überleben mit über 600 TB Rohdaten. Die Kombination von WGS und RNA-Seq sollte Aufschluss darüber geben, welche Mutationen transkribiert und exprimiert werden und welche genetischen Veränderungen die Genregulation und die zellulären Prozesse beeinflussen. Zur Nutzung dieser Daten müssen folgende Herausforderungen bewältigt werden: i) somatische Varianten von WGS müssen ohne passende Kontrollgewebe identifiziert werden, ii) die Varianten müssen interpretiert werden und iii) kausale genetische und Genexpressionsänderungen müssen von lediglich korrelierten Änderungen unterschieden werden.

Die Arbeiten zur Erreichung dieser Ziele gliederten sich in drei Teilprojekte und Ziele:

SP1 Varianten-Identifikation und Varianten-Effektvorhersage (Matthias Heinig)

- 1.1 Deep Learning für die Identifikation somatischer Varianten ohne Normalreferenz
- 1.2 Deep Learning und Populationsgenetik für die Vorhersage von Varianteneffekten

SP2 Driver-Identifikation aus RNA-Sequenz und WGS (Julien Gagneur)

- 2.1 Expressions- und Spleißausreißer in 5k-RNAs-seq-Datensätzen
- 2.2 Priorisierung von Cis-regulatorischen Varianten
- 2.3 Priorisierung von Kandidatengenen für die Metaanalyse über die Proben hinweg

SP3 Krebsdiagnose und Subtypidentifizierung (Stephan Hutter)

- 3.1: Datenauswahl, Vorverarbeitung und Bereitstellung für SP1 und SP2
- 3.2: Subtypidentifizierung und -diagnose
- 3.3: Integrierung molekularer Profile in die Risikostratifizierung des Patienten und die Vorhersage des Behandlungsergebnisses

Es gab keine Änderungen in der Planung und im Ablauf des Vorhabens.

4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Matthias Heinig hat Methoden entwickelt, um die Konsequenzen von Sequenzvariationen auf die Bindung von Transkriptionsfaktoren vorherzusagen und deren Nützlichkeit in Kombination mit Genexpressions- daten (eQTL) gezeigt. Julien Gagneur hat den Nachweis von abnormaler Genexpression zur Diagnose bei seltenen Krankheiten vorangetrieben. Zu

Beginn dieses Projekts hatte Julien Gagneur bereits Pionierarbeit bei der Erkennung von abnormaler Genexpression zur Diagnose bei seltenen Krankheiten geleistet. Insbesondere hat sein Forschungsteam bioinformatische Protokolle entwickelt, um drei Arten von abnormaler Genexpression in RNA-seq-Proben von Patienten zu identifizieren: i) abnormale Expressionslevel, ii) abnormales Spleißen und iii) monoallelische Expression (Kremer et al., 2017; Brechtmann et al. AJHG 2018). Ferner hat seine Forschungsgruppe sequenzbasierte KI-Modelle und Software-Frameworks entwickelt, um den Effekt von genomischen Varianten auf die Genregulation und das Spleißen vorherzusagen.
Um das Verständnis der molekularen Grundlagen von Leukämien und Lymphomen zu verbessern, initiierte das MLL das 5.000-Genom-Projekt, das sich derzeit dem Abschluss nähert. Es wurden WGS- und RNAseq-Daten von 5.000 Patienten erzeugt, die ein breites Spektrum hämatologischer Malignome abdecken, von häufigen Entitäten wie akuter myeloischer Leukämie (AML) oder myelodysplastischem Syndrom (MDS) bis hin zu seltenen Krankheiten wie Haarzellenleukämie (HCL). Erste Analysen dieses Datensatzes durch Forscher des MLL und externer Kooperationspartner konzentrierten sich auf die Charakterisierung der molekularen Profile von Panels bekannter Leukämiegene unter Verwendung von standardmäßigen somatischen Varianten-Callern und manueller Varianten-Interpretation. Diese ersten Ergebnisse wurden auf der 60. Jahrestagung der American Society of Hematology (ASH 2018) vorgestellt

5. Wesentliche Ergebnisse

1. Zur Identifizierung von somatischen Mutationen nur mit Hilfe von WGS Daten von Tumorproben wurde die neue Methode DeepSom entwickelt (1.1). DeepSom besteht aus einem Convolutional Neural Network, das mit Hilfe von Daten zu fünf Krebsarten, bei denen zusätzlich auch noch Proben aus Kontrollgeweben vorhanden waren, trainiert und evaluiert. DeepSom erzielt bessere Ergebnisse als alle zuvor verfügbaren Methoden und steht auf github frei zur Verfügung. Zur Vorhersage der Effekte von Mutationen (1.2) wurden verschiedene unüberwachte Verfahren evaluiert, die mit Hilfe von evolutionären Daten trainiert wurden. Zuerst wurde überprüft, ob die EVE-Methode, die auf einem Autoencoder-Modell für Sequenz-Alignments von protein-kodierenden Bereichen auf das ganze Genom übertragen werden kann. In der zweiten Analyse wurden verschiedene Large-Language-Models auf 3'UTR Sequenzen trainiert. Diese zeigten gute Ergebnisse in verschiedenen Anwendungen, wie z.B. Vorhersage von Genexpression. Alle unüberwachten Modelle zeigten auch gute Vorhersagen der Effekte von Mutationen. Allerdings sind Vorhersagen von überwachten Modellen oder Sequenz-Konservierung nach wie vor besser.
2. Zur Identifizierung von Ausreißern in Genexpression und Splicing (2.1) wurden Methoden basierend auf Autoencodern eingesetzt. Die dadurch entdeckten Gene waren stark für Onkogene angereichert. Einzelne Gene wie LRP1B eigneten sich als Biomarker, um neue Subtypen von Hairy Cell Leukämien zu identifizieren. AbSplice und AbExp wurde zur Vorhersage von Genvarianten verwendet, die falsches Spleißen und abnormale Genexpression verursachen (2.2.) und vermehrt in tumor-relevanten Genen vorkommen. Auf Basis all dieser Ergebnisse wurde ein neues Random Forest Model entwickelt, das neue Krebsgene zuverlässig priorisieren (2.3) kann.
3. Für die Klassifizierung der WHO2022-Subtypen wurden überwachte Verfahren verwendet, die hohe Genauigkeit erreichten (3.1). Zur Identifizierung neuer Subtypen wurden unüberwachte Analysen durchgeführt, die der WHO2022 AML-Klassifikation sehr nahe kommen. Damit könnte man AML-Subtypen allein auf der Grundlage von Sequenzierungsdaten definieren. Allerdings zeigte sich dass, weder die WHO2022 noch die neu identifizierten Subgruppen eine Verbesserungen für die Risikostratifizierung liefern (3.2).

6. Zusammenarbeit mit anderen Stellen.

Primär: Stephan Hutter (Münchner Leukämie Labor), Julien Gagneur (Technische Universität München) und Matthias Heinig (Helmholtz Munich). Des Weiteren: Roland Rad (Technische Universität München), Marc Seifert (Universitätsklinikum Düsseldorf), Christopher C. Oakes (The Ohio State University, USA) und Piers Blombery (University of Melbourne, Australien).

II. Eingehende Darstellung

1. Verwendung der Zuwendung und erzielte wissenschaftliche Ergebnissen

Der wissenschaftliche Bericht umfasst sämtliche Teilprojekte, die in enger Zusammenarbeit zwischen den beteiligten Gruppen von Matthias Heinig, Julien Gagneur und Stephan Hutter bearbeitet wurden. Zu Beginn eines jeden Abschnitts werden die Beiträge der jeweiligen Gruppen erläutert.

SP1 Varianten-Identifikation und Varianten-Effektvorhersage (Matthias Heinig)

Teilprojekt 1 wurde federführend in der Gruppe von Matthias Heinig bearbeitet.

AP1.1 Deep Learning für die Identifikation somatischer Varianten ohne Normalreferenz

Somatische Mutationen werden normalerweise durch die Analyse der DNA-Sequenz einer Tumorprobe zusammen mit einem passenden Normalgewebe festgestellt. Ein passendes Normalgewebe ist jedoch nicht immer verfügbar, zum Beispiel bei retrospektiven Analysen oder in diagnostischen Umgebungen. Für solche Fälle müssen Werkzeuge zur Erkennung somatischer Varianten nur anhand von Tumorgewebe entwickelt werden. Bisher vorgeschlagene Ansätze zeigen bei der Analyse von Ganzgenomsequenzierungsproben (WGS) eine geringere Leistung.

Im VALE Projekt entwickelten wir DeepSom, einen Ansatz, der auf einem Convolutional Neural Network basiert, unde es ermöglicht, somatische Einzelnukleotidpolymorphismen sowie kleiner Insertionen und Deletionen in Tumor-WGS-Proben ohne passendes Normalgewebe zu entdecken. Wir validierten DeepSom, indem wir seine Leistung anhand von fünf verschiedenen Krebsdatensätzen (Tabelle 1) mit passenden Normalgeweben evaluieren (Abb. 1). Wir zeigten auch, dass DeepSom bei WGS-Proben die zuvor vorgeschlagenen Methoden zur Erkennung somatischer Varianten nur anhand von Tumorgewebe in der Leistung übertrifft.

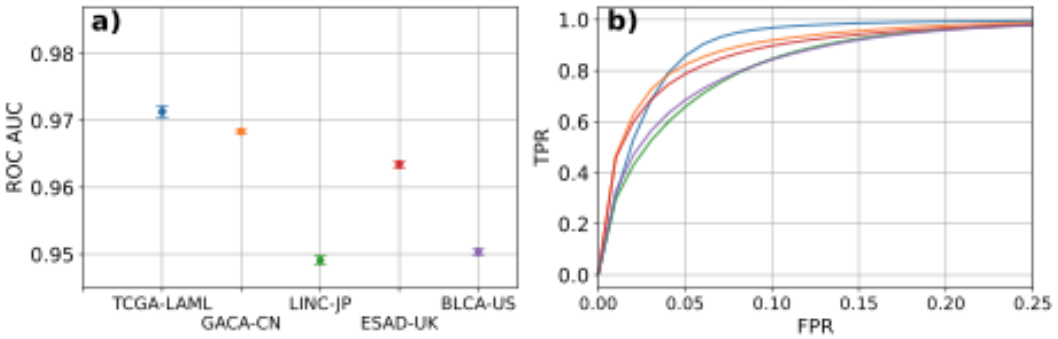


Abb. 1: *Klassifikationsgüte des Variant calling Modells*. Panel a) zeigt die Fläche unter der receiver operator characteristic (ROC) Kurve für verschiedene Initialisierungen des Model-Trainings (Mittelwert und Standard- abweichung) auf der y-Achse und den Datensatz auf der x-Achse. Panel b) zeigt exemplarisch eine ROC Kurve pro Datensatz.

DeepSom ist als GitHub-Repository unter https://github.com/heiniglab/DeepSom verfügbar.

Tabelle 1: *Übersicht über die Trainingsdaten*.

Dataset	Tumor WGS samples	Somatic SNPs per sample	Somatic SNPs in gnomAD v. 3.1.2	Average coverage
TCGA-LAML	47	393	30%	35x
GACA-CN	25	10230	18%	37x
LINC-JP	28	8529	13%	54x
ESAD-UK	41	53212	13%	67x
BLCA-US	23	17998	14%	36x

Die Ergebnisse zum Meilenstein AP1.1.1 Deep-Learning-Modell für die Identifizierung von somatischen Varianten wurden publiziert:
Vilov S, Heinig M. DeepSom: a CNN-based approach to somatic variant calling in WGS samples without a matched normal. Bioinformatics. 2023;39(1):btac828. doi: 10.1093/bioinformatics/btac828.
Da die Anwendung von DeepSom sehr rechenintensiv ist, wurde zunächst ein Test durchgeführt, um zu überprüfen, ob die Verwendung von durch DeepSom identifizierten Varianten zu einer Verbesserung bei der Risikostratifizierung von Patienten führt, welche das schlussendliche Ziel des Projektes darstellt. Dazu wurden die ELN-2022 Klassen für 100 zufällig ausgewählte AML-Patienten bestimmt. Die ELN-Klassifizierung ist in der klinischen Praxis weit verbreitet und verwendet die Diagnosedaten des Patienten, um jedem Patienten eine der drei Prognosekategorien zuzuordnen: „günstig“, „mittel“ oder „ungünstig“. Im Vergleich zum vorherigen System ELN-2017 bezieht ELN-2022 einen größeren Anteil von Punktmutationen in die Risikoklassifizierung ein. Wir wendeten ELN-2022 zunächst unter Verwendung der Mutationen an, die mit unserer ursprünglichen konventionellen Filterpipeline identifiziert wurden. Anschließend verglichen wir diese mit den ELN-2022 Kategorien, die man erhält, wenn die Genomdaten stattdessen mit DeepSom analysiert werden.
DeepSom liefert für jede Variante einen kontinuierlichen Score, dass die Variante somatisch ist. Für die praktische Anwendung benötigt man einen bestimmten Schwellenwert für den DeepSom-Score. Alle Varianten mit Scores über diesem Schwellenwert werden dann als somatisch betrachtet. Der DeepSom-Schwellenwert wurde auf 0,99 festgelegt, da dies eine gute Übereinstimmung zwischen den ursprünglichen und den DeepSom-basierten ELN-2022-Scores ergibt (etwa 78 %, siehe Abb. 2b). Darüber hinaus filtert DeepSom bei diesem Schwellenwert eine akzeptable Anzahl wahrscheinlicher somatischer Varianten (falsch-negative) heraus, was zu nur 4 Proben führt, die aufgrund des Fehlens diagnostischer Marker nicht mit ELN-2022 klassifiziert werden können (Abb. 2a).
Die Verwendung von DeepSom führt zu einer Umverteilung der ELN-2022-Kategorien unter den betrachteten Patienten (Abb. 2c): Bei 17 Proben ändert sich die Kategorie von „ungünstig“ zu „mittelschwer“. Wie bereits erwähnt, können 4 Proben (3 aus der Kategorie „Ungünstig“ und 1 aus der Kategorie „Mittel“) nicht klassifiziert werden, da die ELN-2022-Diagnosemarker wahrscheinlich fälschlicherweise von DeepSom entfernt wurden.
Anschließend trainierten wir ein Cox-Proportional-Hazards-Modell unter Verwendung der ursprünglichen und neu berechneten ELN-2022-Werte. In beiden Fällen fügten wir auch Alter und Geschlecht als Modellmerkmale hinzu. Das Modell wurde mit einer 10-fachen Kreuzvalidierung bewertet, die 5-mal wiederholt wurde. Der resultierende Harrel's c-index für

alle Läufe der Kreuzvalidierung ist in Abb. 2d dargestellt. Der durchschnittliche c-Index beträgt 0,536±0,109 und 0,555±0,104 für das ELN-2022 und das DeepSom-basierte ELN-2022 Modell. Der Unterschied zwischen den Modellen ist statistisch nicht signifikant (p-value=0,3, gepaarter t-Test).

Zusammenfassend zeigt unsere Analyse keinen Hinweis darauf, dass die Ersetzung der herkömmlichen Filterpipeline durch DeepSom zu einer deutlich verbesserten Überlebensvorhersage führt. Dies ist wahrscheinlich auf den sehr begrenzten prognostischen Effekt der meisten der betrachteten molekularen Aberrationen zurückzuführen. Da die Anwendung von DeepSom zu keiner Verbesserung der Risiko-Stratifizierung führte, wurde auf die rechenintensive Anwendung auf der gesamten MLL Kohorte verzichtet.

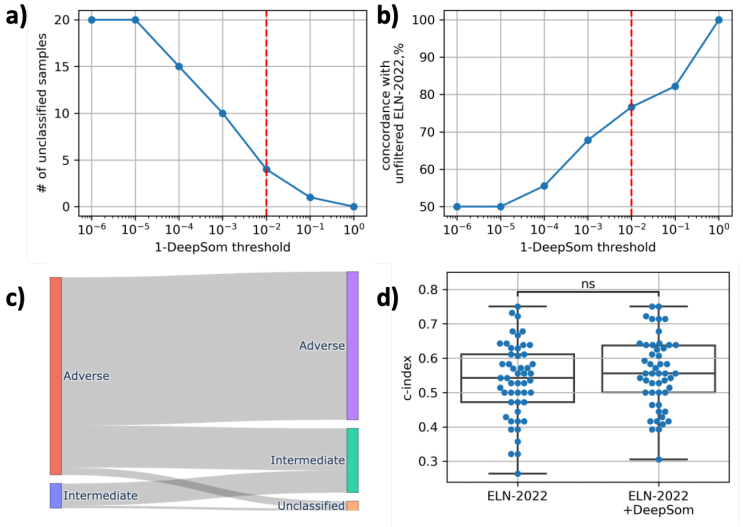


Abb. 2 *Vergleich der Überlebensanalysen von AML Patienten zwischen DeepSom und der bisherigen Filterstrategie*. (a) Anzahl der nicht klassifizierten Stichproben als Funktion des DeepSom-Schwellenwerts, (b) Übereinstimmung zwischen den ursprünglichen ELN-2022-Kategorien und den ELN-2022-Kategorien nach der DeepSom-Filterung, (c) Umverteilung der ELN-2022-Risikokategorien als Ergebnis der DeepSom-Filterung bei einem Schwellenwert von 0,99, (d) Harrel's c-Index für das Cox-PH-Modell, das anhand der ursprünglichen ELN-2022-Zuordnung und nach der DeepSom-Filterung trainiert wurde.

AP1.2 Deep Learning und Populationsgenetik für die Vorhersage von Varianteneffekten

Ursprunglich war geplant die Effekte von Varianten mit Hilfe eines Populationsgenetischen Ansatzes (CADD) vorherzusagen. In der Zwischenzeit sind jedoch Ergebnisse bekannt geworden, die zeigen, dass evolutionäre Vergleiche über eine große Anzahl von Spezies eine mindestens ebenso gute Vorhersage ermöglichen (Frazer el al., Nature 599.7883 (2021)). Daher evaluierten wir zwei Ansätze, die auf selbstüberwachtem Lernen (self supervised learning) basieren.

Das erste Model nennt sich EVE und baut auf einen evolutionären Autoencoder auf. EVE wurde bislang jedoch nur auf Protein-kodierenden Sequenzen getestet. Daher haben wir das Model zunächst so modifiziert, dass es statt Alignments von Aminosäure-Sequenzen Alignments von DNA-Sequenzen verarbeiten kann. EVE prozessiert Sequenz-Alignments von 500bp Länge, die 600 verschiedene Spezies umfassen (Armstrong et al. Nature 587.7833 (2020)). Der Autoencoder wird über die Sequenzen der verschiedenen Spezies trainiert (jeder Datenpunkt repräsentiert eine Sequenz einer Spezies). Der evolutionäre Index schätzt für jede Position in der Sequenz, wie schwer es ist, die Sequenz zu rekonstruieren. Wenn eine Variante pathogen ist, sind mutierte Sequenzen sehr selten und daher schwieriger zu rekonstruieren für den Autoencoder. Der evolutionäre Index wird für jede

Position und jede mögliche Substitution berechnet. Die Verteilung dieser 500 x 3 = 1500 evolutionären Indizes wird durch ein Gauss'sches Mischmodel mit zwei Komponenten geschätzt. Die Komponente mit den höheren evolutionären Indizes repräsentiert pathogene Varianten. Das Mischmodel kann also dafür verwendet werden jeder möglichen Mutation die Wahrscheinlichkeit zu-zuordnen, dass die Mutation pathogen ist. Das angepasste Model wurde anhand von bekannten pathogenen Varianten aus der ClinVar Datenbank getestet. Tabelle 2 zeigt, dass EVE auf Protein-kodierenden Varianten besser pathogene Varianten besser identifiziert als der weit verbreitete PhyloP score. Die Performance auf den nicht-kodierenden Varianten ist Vergleichbar mit der von PhyloP. Im nächsten Schritt werden noch Vergleiche mit anderen Methoden für die nicht-kodierenden Varianten durchgeführt.

Tabelle 2. Vorhersagegüte (ROC AUC) auf bekannten pathogenen Varianten aus der ClinVar Datenbank.

	EVE 600way				PhyloP 100way
	no reweighting	diversity* (theta=0.1)	distance to human*	only mammals	
coding	0.916/0.961	0.962/0.969	0.913/0.955	0.916/0.961	0.925
non-coding		0.96/0.968			0.976

Der zweite selbstüberwachte Ansatz basiert auf sogenannte "foundation models", welche einen weiteren sehr vielversprechenden Ansatz zur Verarbeitung von evolutionären Informationen in Genomsequenzen bieten. Diese Modelle sind eine Art von Large Language Model, wie sie auch bei der Verarbeitung von natürlicher Sprache eingesetzt werden (z.B. ChatGPT). Im VALE Projekt haben wir evaluiert, ob sich Modelle wie DNABERT und Nucleotide Transformer auch für die Vorhersage der Effekte von Sequenzvarianten eignen. Zuvor wurden diese Modelle häufig auf ganzen Genomen trainiert und bewertet. Dabei wird die Aufteilung des Genoms in verschiedene funktionelle Regionen vernachlässigt. Um eine Einschätzung der Qualität dieser Modelle für die Vorhersage von Effekten von Sequenzvarianten zu erhalten, haben wir uns auf 3'UTR-Regionen konzentriert, da diese gut definierte regulatorische Regionen darstellen. Ausserdem gibt es eine Vielzahl von Datensätzen, welche die Funktion der Seqeunzen experimentell belegen. Unsere Bewertung der Modelle umfasst Vorhersagen, die für die RNA-Biologie spezifisch sind. Dies umfasst die Erkennung von Bindungsmotiven von RNA-bindenden Proteinen, die Erkennung funktioneller genetischer Varianten, die Vorhersage von Expressionsniveaus in massiv-parallelen Reporter-Assays und die Abschätzung der mRNA-Halbwertszeit. Bemerkenswerterweise zeigen Modelle, die speziell auf 3'UTR-Sequenzen trainiert wurden, im Vergleich zu den etablierten genomweiten Grundmodellen in drei der vier Analysen eine überlegene Leistung. Diese Ergebnisse unterstreichen, wie wichtig es ist, die Aufteilung des Genoms in funktionale Regionen bei der Ausbildung und Bewertung von Basismodellen zu berücksichtigen.

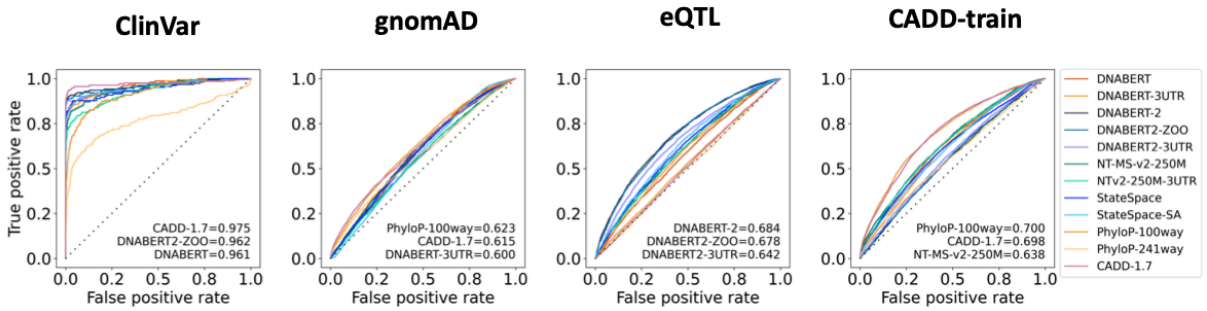


Abb. 3. *ROC Kurven für die Vorhersage von funktionellen Varianten in 3'UTR*. Die drei besten Modelle sind in der Legende der Paneele gezeigt. Die Überschriften der Paneele geben an, welche Art von funktionalen Varianten vorhergesagt wurden. ClinVar bezeichnet kausale Varianten für monogene Erkrankungen, gnomAD bezeichnet seltene Varianten in der gnomAD Datenbank, eQTL bezieht sich auf regulatorische Varianten, CADD-train bezeichnet den Trainingsdatensatz der CADD Methode.

In Bezug auf die Qualität der Vorhersage der Effekte von Sequenzvarianten in nicht-kodierenden Bereichen (Abb. 3) lassen sich die Ergebnisse wie folgt zusammenfassen. Bislang können die selbstüberwachte Ansätze, die in diesem Projekt untersucht wurden noch nicht die Qualität der Vorhersagen mit Hilfe von Sequenzkonservierung (PhyloP) oder von Modellen, die mittels überwachtem Lernen trainiert wurden (CADD), erreichen. Nichtsdestotrotz sind die Ergebnisse bezüglich der Vorhersage anderer regulatorischer Eigenschaften hoch interessant und wurden zur Veröffentlichung eingereicht und stehen als Preprint zur Verfügung:

Vilov S, Heinig M. Investigating the performance of foundation models on human 3'UTR sequences. bioRxiv 2024; doi: 10.1101/2024.02.09.579631

Wie geplant wurden die Effekte aller Varianten in der gesamten Patientenkohorten mit dem Foundation Model annotiert. Da das Modell spezifisch für regulatorische Varianten in der 3'UTR ist, wurde die Qualität dieser Vorhersagen auch spezifisch in 3'UTR Regionen evaluiert. Die Testdaten wurden auf die gleiche Weise wie die Trainingsdaten vorbereitet, wobei MLL 5K-Varianten verwendet wurden, die sich mit ClinVar-, gnomAD-, eQTL- und CADD-Kohorten überschneiden. Insgesamt wurde die folgende Anzahl unterschiedlicher Mutationen mit wahrscheinlicher funktioneller Auswirkung entdeckt: 2 ClinVar-Varianten mit (wahrscheinlich) pathogenen Annotationen (rs121965020, rs772838513, zuvor nicht mit Leukämie assoziiert), 4.541 Mutationen, die sich mit dem positiven CADD-Satz überschneiden, 58.127 extrem seltene Mutationen (solche die nur in einem einzigen Individuum in der gnomAD vorkommen), und 1.078 Mutationen aus dem glaubwürdigen eQTL Susie-Satz mit einem p-Wert<1e-12. Um die gnomAD- und eQTL-Daten zu testen, wurde jeder positive Satz mit 10.000 wahrscheinlich gutartigen 3'UTR-spezifischen SNPs (gnomAD-Population AF>5%) ergänzt, die in der MLL 5K-Kohorte entdeckt wurden. Für die Trainingsdaten wurde ein separater Satz negativer Beispiele für die CADD-Kohorte verwendet, der aus 6.064 MLL5K 3'UTR-spezifischen SNPs bestand, die sich mit dem negativen CADD-Satz überschnitten.

Die Leistung der MLP-Modelle, sowie einfacherer Konservierungsmethoden, wird auf den Trainings- und Testdaten wird in Abb. 4 und in Tabelle 3 gegenübergestellt. Wie schon zuvor beschrieben ist die Vorhersagequalität im allgemeinen für die gnomAD und CADD Datensätzen bei den Konservierungs-basierten, sowie überwachten Modellen am besten. Bei der Evaluierung auf den MLL5K Varianten zeigt sich, dass die neuen Vorhersagen basierend auf DNABERT Modellen, die für regulatorische Varianten (eQTL) trainiert wurden eine Verbesserung gegenüber den anderen Methoden liefern. Insgesamt, ist festzustellen, dass die Vorhersage auf regulatorischen Varianten am besten funktioniert. Diese Art von Varianten

ist für jedoch für klinische Experten nur mit Hilfe von zusätzlichen experimentellen Daten einzuschätzen. Da solche Daten nicht verfügbar waren, wurde konnte keine Bewertung durch klinische Experten durchgeführt werden. Überwachte und Konservierungsmethoden für die Vorhersage von Krankheitsrelevanz (CADD und gnomAD) bessere Leistungen erbringen. Daher ist es zu bevorzugen, zunächst noch diese klassischen Methoden für die Priorisierung von kausalen Varianten zu verwenden.

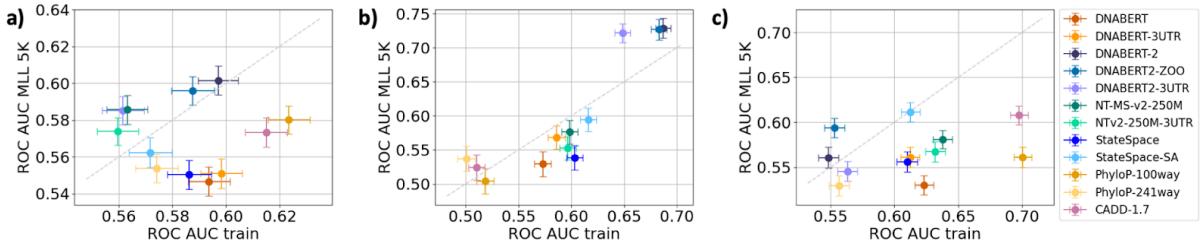


Abb. 4. *ROC-AUC-Scores für MLP- und Konservierungs-basierte Modelle, die anhand der Training - und Testdaten (MLL 5K) für die gnomAD-, eQTL- und CADD-Kohorten ermittelt wurden*. Die Fehlerbalken zeigen Bootstrap-geschätzte 95%-Konfidenzintervalle an.

Model	gnomAD		eQTL-susie		CADD	
	train	test (MLL5K)	train	test (MLL5K)	train	test (MLL5K)
DNABERT	0.594±0.008	0.547±0.008	0.573±0.008	0.529±0.018	0.623±0.008	0.530±0.011
DNABERT-3UTR	0.598±0.008	0.551±0.008	0.586±0.008	0.568±0.017	0.612±0.007	0.561±0.011
DNABERT-2	0.597±0.007	0.601±0.008	0.687±0.007	**0.729±0.014**	0.549±0.008	0.561±0.011
DNABERT2-ZOO	0.588±0.008	0.596±0.008	0.683±0.007	**0.727±0.015**	0.553±0.008	0.594±0.011
DNABERT2-3UTR	0.562±0.008	0.585±0.008	0.648±0.008	0.721±0.014	0.563±0.008	0.545±0.011
NT-MS-v2-250M	0.563±0.008	0.586±0.008	0.599±0.008	0.576±0.018	0.638±0.008	0.581±0.010
NTv2-250M-3UTR	0.560±0.008	0.574±0.007	0.597±0.008	0.553±0.017	0.632±0.008	0.567±0.011
StateSpace	0.586±0.008	0.550±0.008	0.603±0.008	0.539±0.018	0.610±0.008	0.556±0.011
StateSpace-SA	0.572±0.008	0.562±0.008	0.616±0.008	0.594±0.017	0.613±0.007	0.611±0.011
PhyloP-100way	**0.623±0.008**	0.580±0.008	0.519±0.008	0.504±0.019	**0.700±0.007**	0.561±0.011
PhyloP-241way	0.574±0.008	0.554±0.008	0.501±0.008	0.537±0.018	0.557±0.008	0.529±0.011
CADD-1.7	**0.615±0.008**	0.573±0.008	0.510±0.008	0.524±0.019	**0.698±0.007**	0.608±0.011

Tabelle 3. ROC-AUC-Werte für MLP- und Konservierungs-basierte Modelle, die anhand der Training - und Testdaten (MLL 5K) für die gnomAD-, eQTL- und CADD-Kohorten ermittelt wurden. Das Bootstrap-geschätzte 95%-Konfidenzintervall ist angegeben.

SP2 Driver-Identifikation aus RNA-Sequenz und WGS (Julien Gagneur)

Wir beschreiben hier unsere Ergebnisse gemäß den 3 Zielen des Teilprojekts SP2. Dieses Teilprojekt wurde unter der Federführung von Julien Gagneur bearbeitet. Personal, das aus Teilprojekt 1 finanziert wurde, leistete wichtige Beiträge zu den Analysen der Sequenzvarianten. Die folgenden Ergebnisse sind Teil einer Veröffentlichung (Cao, X. *et al.* Analysis of 3760 hematologic malignancies reveals rare transcriptomic aberrations of driver genes. *Genome Med.* 2024. **16**(70)), in der wir WGS-und RNA-seq-Daten von 3.760 Tumorproben untersucht haben, die aus insgesamt 24 verschiedenen Leukämie- und Lymphom-Typen stammen (Abb. 4). Dies ist die größte Sammlung von Blutproben mit pathologischen Befunden, die auch seltene Krankheitsbilder wie die Haarzellleukämie-Variante (HCL-V) und die chronisch lymphoproliferative Störung der natürlichen Killerzellen umfasst.

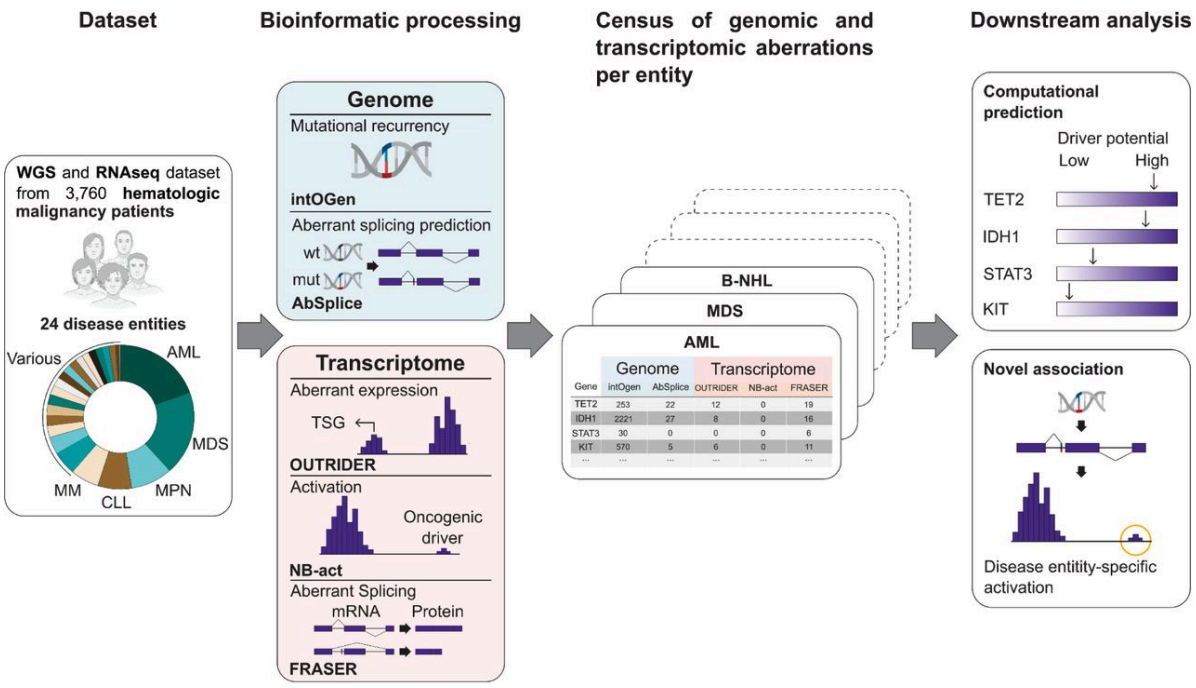


Abb. 4. Schema der Hauptstudie. Datensatz: WGS und RNA-seq Daten von 3.760 hämatologischen Malignomen, die 24 verschiedene Krankheits Subtypen umfassen. Bioinformatische Verarbeitung: intOGen detektiert wiederkehrende Mutationsmuster (IntOgen) in genomischen Daten und AbSplice prognostiziert Varianten, die abnormales Spleißen verursachen. OUTRIDER detektiert abnormale Expression häufig exprimierter Gene [REF], NB-act detektier die Überexpression selten exprimierter Gene ab (für die Studie entwickelt), und FRASER identifiziert abnormales Spleißen (Scheller I. et al. Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index. Am J Hum Genet. 2023; 110(12)). Zensus: Eine einzigartige Sammlung von genomischen und transkriptomischen Auffälligkeiten in 24 Subtypen hämatologischer Malignome. Anschließende Analyse: Vorhersage von Krebsgenen und deren spezifische Häufung in verschiedenen Leukämie-Subtypen.

AP2.1 Expressions- und Spleißausreißer in den 5k-RNAs-seq-Datensätzen

Expressionsaureißer mit OUTRIDER

Die MLL-Daten wurden mit Hilfe der Software DROP (Yépez V.A. et al. Detection of aberrant gene expression events in RNA sequencing data. Nat Protoc. 2021; (16)) prozessiert um mögliche statistisch abnormale Spleiß- und Expressionsereignisse zu detektieren. Dazu musste das Softwaremodul OUTRIDER – ein Bestandteil von DROP- in ein Python Software Modul basierend auf TENSORFLOW umgeschrieben werden (pyOUTRIDER). Dadurch konnte OUTRIDER die 4.000 Proben innerhalb eines Tages analysieren.

Nach eingehenden Qualitätskontrollen der Daten haben wir die Studie mit ~3.800 aus 4.000 verfügbaren Proben fortgesetzt. Mit Hilfe von pyOUTRIDER in Kombination mit DROP konnten wir ca. 34k Expressionsausreißer in den 3.800 Proben identifizieren.

Um zu überprüfen, ob diese Ausreißer mit Krebs assoziiert sind, haben wir ihre Häufung unter bekannten Krebsgenen hämatologischer Erkrankungen, die wir aus der CGC Datenbank (42) übernommen haben, gemessen. Wir fanden eine starke Häufung für Tumorsuppressorgene unter den Unterexpressionsausreißern und für Onkogene unter den Überexpressionsausreißern (Abb. 5). Bemerkenswert ist, dass die Gene, die in mehr als fünf Proben als Ausreißer gefunden wurden, die höchste Häufung aufwiesen, was darauf hindeutet, dass Gene, die häufig als Ausreißer bezeichnet werden, mit größerer Wahrscheinlichkeit onkogen sind.

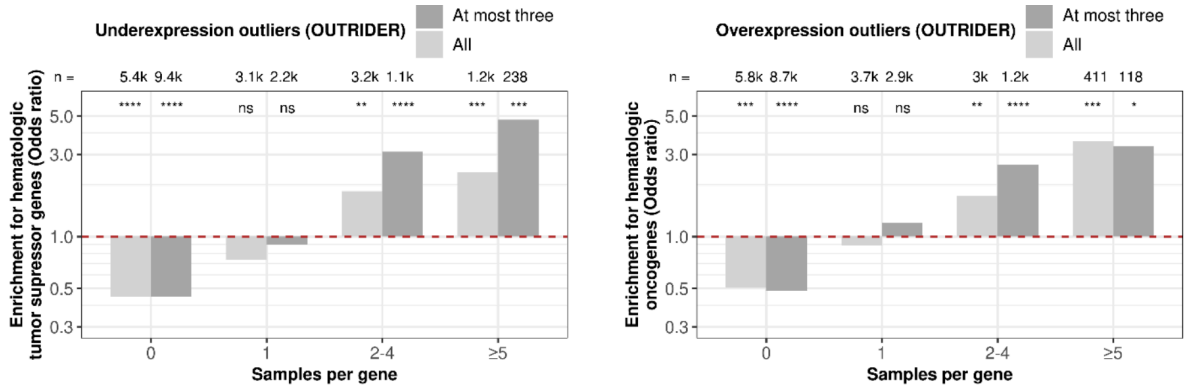


Abb. 5 Expressionsausreißer finden sich gehäuft in Onkogenen hämatologischer Erkrankungen. (Links) Häufung von aus CGC bekannten Tumorsuppressorgenen hämatologischer Erkrankungen unter allen Genen, die von OUTRIDER als unterexprimiert eingestuft wurden, sowie unter den TOP drei signifikanten Genen pro Probe, die als Unterexpressionsausreißer eingestuft wurden. Die Gene wurden nach der Anzahl der Proben aufgeteilt, in denen das Gen als Ausreißer identifiziert wurde. Die Anzahl der Gene und die nominalen Signifikanzen aus dem Fisher-Test sind oben in den Balken angegeben (ns: nicht signifikant; *: P ≤ 0.05; **: P ≤ 0.01; ***: P ≤ 0.001; ****: P ≤ 0.0001). (Rechts) Wie links) für CGC Onkogene hämatologischer Erkrankungen unter OUTRIDER-Überexpressionsausreißern.

Aktivierungsausreißer mit NB-act

Darüber hinaus haben wir unsere Methode zur Identifikation von Genen mit stark abweichender Expression (expression outlier) erweitert, um die Aktivierung von Genen, einschließlich Onkogenen, die normalerweise nicht exprimiert werden, besser zu erfassen.
Die Methode NB-act (Negative-Binomial Aktivierung), liefert P-Werte für die beobachtete Zahl von RNA-Fragmenten (read pair) für jedes Gen in jeder Probe unter der Nullhypothese, dass das Gen in der Probe nicht exprimiert wird. Konkret berechnet NB-act die Wahrscheinlichkeit, mehr als eine bestimmte Anzahl von Fragmenten eines Gens in einer bestimmten Probe zu beobachten, unter der Annahme einer negativen Binomialverteilung mit einer angenommenen Grundexpression von 1 FPKM und einem Streuungsparameter von 0,02.
Wir haben unsere Methode bei 6.017 selten exprimierten und proteinkodierenden Gene angewendet, die von OUTRIDER nicht berücksichtigt wurden. NB-act identifizierte 10.263 Aktivierungsausreißer in 1.623 Genen (mit einem 75%-Quantil von 2 pro Probe). Wir beobachteten eine signifikante Häufung von CGC-Onkogenen hämatologischer Erkrankungen unter allen Aktivierungsausreißern (Abb. 6). Auch hier erhöhte die Beschränkung auf höchstens drei Ausreißer pro Probe deren Häufigkeit.

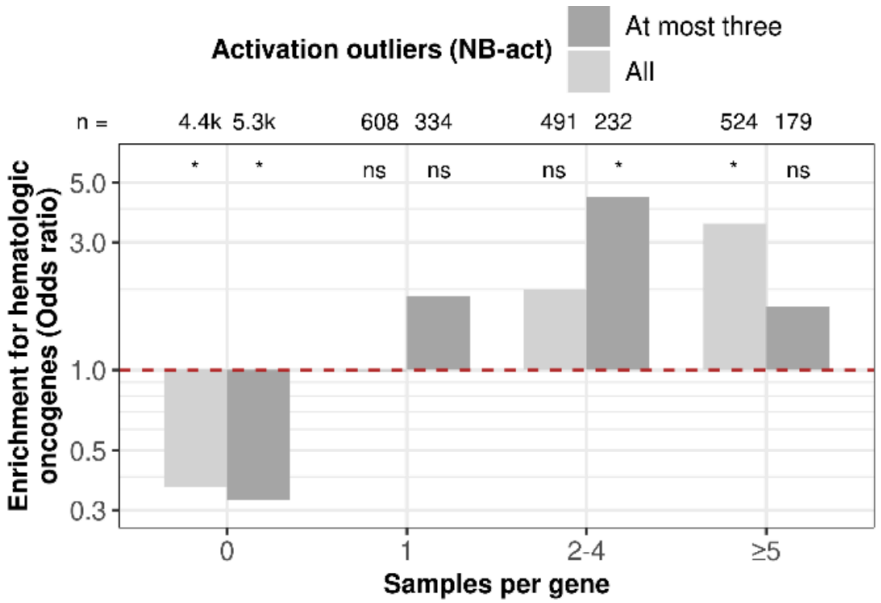


Abb. 6. Aktivierungsausreißer finden sich gehäuft in Onkogenen hämatologischer Erkrankungen. Wie in Abb. 5, jedoch für Gene, deren Aktivierung mit NB-act nachgewiesen wurde.

Zusammenfassend lässt sich sagen, dass mit Hilfe von OUTRIDER als auch von NB-act abnormal exprimierte Gene identifiziert werden können, die stark gehäuft bereits als Krebsgene bekannt sind. Dies ermöglicht die Identifizierung von unterdrückten Tumorsuppressorgenen oder aktivierten Onkogenen in spezifischen Proben.

Spleißausreißer

FRASER ist ähnlich wie OUTRIDER für Expression ein Werkzeug, um abnormales Spleißen zu identifizieren, wobei man Kovariationen herausgerechnet. Dies unterscheidet FRASER von anderen Methoden, die auf dem differentiellen Vergleich verschiedener Gruppen basieren. In einem weiteren, parallel laufenden Projekt wurde eine verbesserte Version von FRASER, FRASER 2.0 (Scheller I. et al. Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index. Am J Hum Genet. 2023; 110(12)), entwickelt, die wir nun auf den gesamten Datensatz angewendet haben.

Wir wendeten FRASER an, um abnormales Spleißereignisse (abweichende Nutzung bestehender oder neuer Spleißstellen) in unseren Proben zu erkennen, die durch Ereignisse wie alternative Exon-Nutzung, Intron-Retention, alternative Donor- oder Akzeptorstellennutzung, Nutzung tiefer intronischer Donor- und Akzeptorstellen oder Kürzungen von Teilen des Transkripts verursacht werden könnten. Wir haben 43.464 Spleißausreißer in 35.410 Spleißausreißer-Ereignissen auf Gen-Ebene bei insgesamt 7.591 Genen in 2.854 Proben festgestellt. Bemerkenswerterweise beobachteten wir eine erhebliche Häufung von aus CGC bekannten Tumorsuppressorgene hämatologischer Erkrankungen unter diesen Spleißausreißern (Abb. 7). Wie bei den Expressionsausreißern erhöhte sich die Häufung, wenn man sich auf höchstens drei Ausreißer pro Probe beschränkte. Außerdem wiesen die Gene, die in mehr als fünf Proben als Spleißausreißer identifiziert wurden, die höchste Häufung auf, was auf ein höheres onkogenes Potenzial schließen lässt. Diese Ergebnisse deuten darauf hin, dass seltene Spleißabweichungen, die aus der RNA-seq Daten ermittelt werden, zur Identifizierung potenzieller Treibergene bei hämatologischen Malignomen beitragen können.

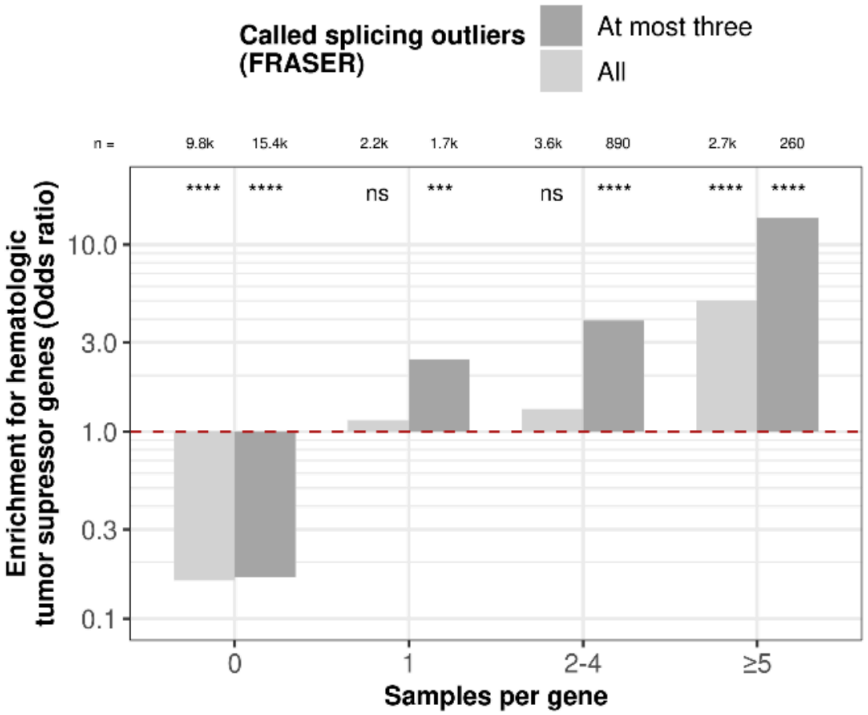


Abb. 7 Abnormales Spleißen finden sich gehäuft in Krebsgenen hämatologischer Erkrankungen. Häufung von aus CGC bekannten Tumorsuppressorgene hämatologischer Erkrankungen unter allen von FRASER identifizierten Genen und unter den drei am höchsten signifikanten Genen pro Probe, die von FRASER als abnormal gespleißt identifiziert wurden. Die Gene sind nach der Anzahl der Proben geordnet, in denen ein Gen als abnormal gespleißt identifiziert wurde. Die Anzahl der Gene und ihre nominale Signifikanz aus dem Fisher-Test sind oberhalb der Balken angegeben.

AP2.2 Priorisierung von cis-regulatorischen Varianten

Um unsere Analyse auf seltene Keimbahn- und somatische Varianten zu beschränken, filterten wir die WGS-Varianten mit strengen Qualitätsfiltern und basierend auf der Verteilung der Allelfrequenz in der Gesamtbevölkerung. Anschließend haben wir die Gene mit Hilfe der Software intOGen (Martínez-Jiménez F et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10)) annotiert, zu denen das positionsbezogene Wiederauftreten von Mutationen in der Genomsequenz (OncodriveCLUSTL), das positionsbezogene Wiederauftreten von Mutationen in der Proteinkonformation (HotMAPS) und die Anreicherung von Mutationen in funktionellen Domänen (smRegions) gehören, drei verschiedene Maße für die Selektionsstärke, die aus synonymen und nicht-synonymen Mutationen abgeleitet werden (CBaSE, MutPanning und dNdScv), und OncodriveFML, eine Methode, die ein Häufung von somatischen Mutationen in Tumoren sowohl in kodierenden als auch in nicht-kodierenden Genomregionen identifiziert.
In einem parallel laufenden Projekt haben wir eine neue Methode, genannt AbSplice, zur Vorhersage von Genvarianten, welche falsches Spleißen zur Folge haben, entwickelt (Wagner N. et al. Aberrant splicing prediction across human tissues. Nat Genet. 2023; 55). Wir haben die Methode auf den gesamten Datensatz angewandt, um Varianten zu annotieren, die mit großer Wahrscheinlichkeit ursächlich für falsches Spleißen sind. Dabei haben wir eine Häufung solcher Varianten in Tumor relevanten Genen gefunden (Abb. 8).

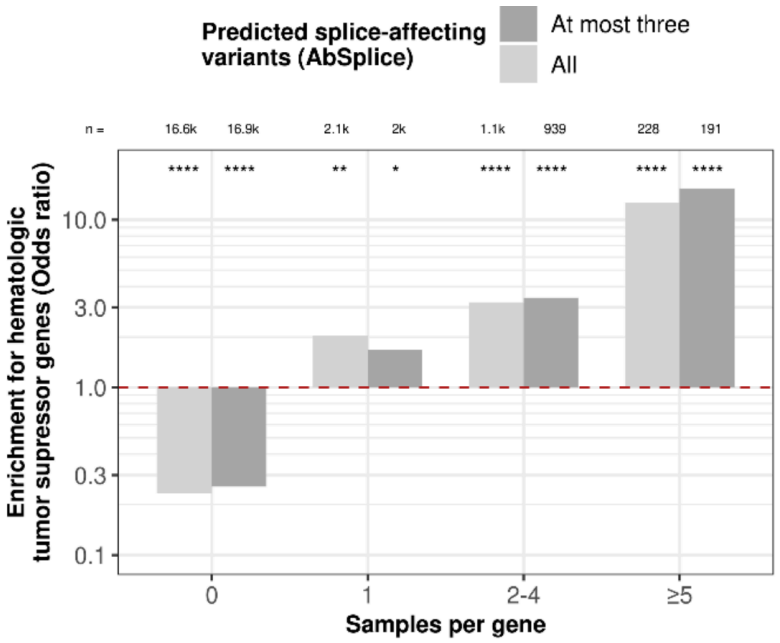


Abb. 8 Varianten, die das Spleißen beeinträchtigen, finden sich gehäuft in Krebsgenen hämatologischer Erkrankungen. Ähnlich wie in Abb. 6, aber für Varianten, die das Spleißen beeinträchtigen und mit AbSplice vorhergesagt wurden.

Darüber hinaus haben wir eine neue Methode namens AbExp entwickelt, um abnormale Genexpression auf Basis seltener Genvarianten vorherzusagen. Hierfür haben wir uns zunächst auf Daten von gesunden Spendern konzentriert, die bessere Voraussetzungen zur Etablierung derartiger Vorhersagemodelle bieten. Ein Vorabdruck unserer Ergebnisse ist online verfügbar unter: https://www.biorxiv.org/content/10.1101/2023.12.04.569414v1. Die Anwendung dieser Methoden auf die MLL-Kohorte ist für eine anschließende Kollaboration geplant.

AP2.3 Priorisierung von Kandidatengenen für die Metaanalyse über die Proben hinweg

Driver gene predictions

Zur Vorhersage von Krebsgenen trainierten wir Modelle auf der Grundlage genomischer und transkriptomischer Merkmale, u.A. mit Hilfe des Outputs der sieben intOGen-Tools, AbSplice, OUTRIDER, NB-act und FRASER. Das Modell wurde sowohl auf dem vollständigen Datensatz als auch auf dem nach 14 Studiengruppen unterteilten Datensätzen trainiert. Als Referenz dienten die 322 bekannten CGC-Krebsgene hämatologischer Krankheiten, die durch 55 zusätzliche, kuratierte hämatologische Panel-Gene ergänzt wurden.

Unter Verwendung des vollständigen Datensatzes stellten wir fest, dass die genomischen und transkriptomischen Merkmale sich bei der Vorhersagewert von Krebsgenen ergänzen (Abb. 9A-B). Insbesondere verbesserte die Integration von AbSplice-Varianteneffektvorhersagen das genombasierte Modell, das mit dem Signal aus den sieben intOGen-Tools trainiert wurde, erheblich. Darüber hinaus stellten wir fest, dass die transkriptomischen Merkmale das Modell weiter deutlich verbesserten (Abb. 9B). Diese Ergebnisse unterstreichen die Relevanz der Einbeziehung abnormaler Expressions und Spleißens zur Vorhersage von Krebsgenen. Von den 100 Genen mit den besten Vorhersagen waren 63 aus der Referenz bekannt, was weit über deren Erwartungswert hinausgeht (Odds Ratio = 106,3, P = 1,6×10-84, Fisher-Test; Abb. 9C). Diese Häufung für bekannte Krebsgene demonstriert die Zuverlässigkeit der Modellvorhersagen. Bei vier Genen, CDNKN1B, EIF3E, HLA-A und IL6ST, handelt es sich um CGC-Krebsgene, die noch nicht im hämatologisch Kontext bekannt sind, was auf eine umfassendere Rolle dieser Gene hinweist.

In Übereinstimmung mit den Ergebnissen des gesamten Datensatzes übertraf unser integratives Modell intOGen oder MutSigCV - ein auf der Mutationshäufigkeit basierendes Krebsgen Vorhersagemodell - in allen Studiengruppen oder war gleichwertig mit diesen. Die studiengruppenspezifischen Modelle ermöglichten weitere Einblicke in die Spezifitäten der Krankheits-subtypen (Abb. 9D). Ein Clustern der 100 bestplatzierten Gene nach den von den Studiengruppen vorhergesagten Wahrscheinlichkeiten ergab verschiedene krankheitsspezifische Spezialfälle (siehe Cao et al.).
Unter den möglichen Krebsgenen identifizierten wir mehrere vielversprechende Gene, deren spezifische Rolle bei hämatologischen Malignomen noch zu klären ist. Unsere Analyse ergab, dass SORL1 in mehreren Studiengruppen, darunter myelodysplastische Neoplasien (Vorläufer der AML), B-Zell-Vorläufer der ALL und eine Studiengruppe mit T-Zell-Non-Hodgkin-Lymphom und T-Zell-akuter lymphoblastischer Leukämie, als mögliches Krebsgen in Frage kommt. In Übereinstimmung mit diesen Beobachtungen wurde festgestellt, dass SORL1 bei AML und ALL auf der Zellmembran der leukämischen Zellen exprimiert und ins Plasma freigesetzt wird, wobei seine Aktivität während der Remission abnimmt (Sakai S. et al. Circulating soluble LR11/SorLA levels are highly increased and ameliorated by chemotherapy in acute leukemias. Clin Chim Acta. 2012;413(19)). EIF4G1 erwies sich als weiterer interessanter Kandidat für AML und lymphoide Krankheitstypen, was durch frühere Analysen unterstützt wird, die darauf hindeuten, dass EIF4G1 als nachgeschaltetes Ziel von MYCN, einem bekannten Onkogen im Neuroblastom, am Zellüberleben bei AML beteiligt ist (Peramangalam PS et al. MYCN Regulates Cell Survival Via EIF4G1 in Acute Myeloid Leukemia. Blood. 2022;140(Supplement 1)). Weitere interessante Beispiele für Kandidaten werden in der Veröffentlichung beschrieben. Insgesamt zeigen unsere Vorhersagen auf der Grundlage von 3.760 Blutproben vielversprechende Krebsgenkandidaten in hämatologischen Malignomen auf.

Darüber hinaus haben wir eine neue statistische Methode zur Analyse von Daten aus Transposon-Screens entwickelt, die unsere bestehenden Analysen der Patientenproben ergänzen und unseren Algorithmus zur Vorhersage tumortreibender Gene verbessern könnte Die Methode wurde in Nucleic Acids Research (Bredthauer C. et al. Transmicron: accurate prediction of insertion probabilities improves detection of cancer driver genes from transposon mutagenesis screens. Nucleic Acids Res. 2023; 51(4)) veröffentlicht und findet Krebsgene mit höherer Genauigkeit als bisherige Methoden für Transposon-Screen-Analysen. Für die Zukunft planen wir die Integration von Genen, die durch Transposon-Screens entdeckt wurden, mit solchen, die bei Patienten aus den MLL-Datensätzen gefunden wurden, um unser Verständnis von Krebsgenen und onkologischen Signalwegen in Leukämie weiter zu vertiefen.

LRP1B activation as a new biomarker for HCL-V
Zusätzlich zu unserer globalen Vorhersage von Krebsgenen führten wir eine detaillierte Untersuchung in den einzelnen Krankheitssubtypen durch und untersuchten sie auf abnormale Expression, abnormes Spleißen und auf Varianten, die abnormes Spleißen verursachen können. Insgesamt fanden wir 2.716 signifikante Zusammenhänge zwischen 11.273 Genen und 24 Krankheitsentitäten. Wenn wir uns auf Aktivierungsausreißer und annotierte Krebsgene beschränkten, fanden wir 43 Zusammenhänge zwischen 37 CGC-Krebstreibergenen und 12 Krankheitsentitäten (Abb. 10A). Einige Assoziationen wurden bereits in der Literatur beschrieben (Einzelheiten siehe Originalveröffentlichungen), was die Ergebnisse untermauert.

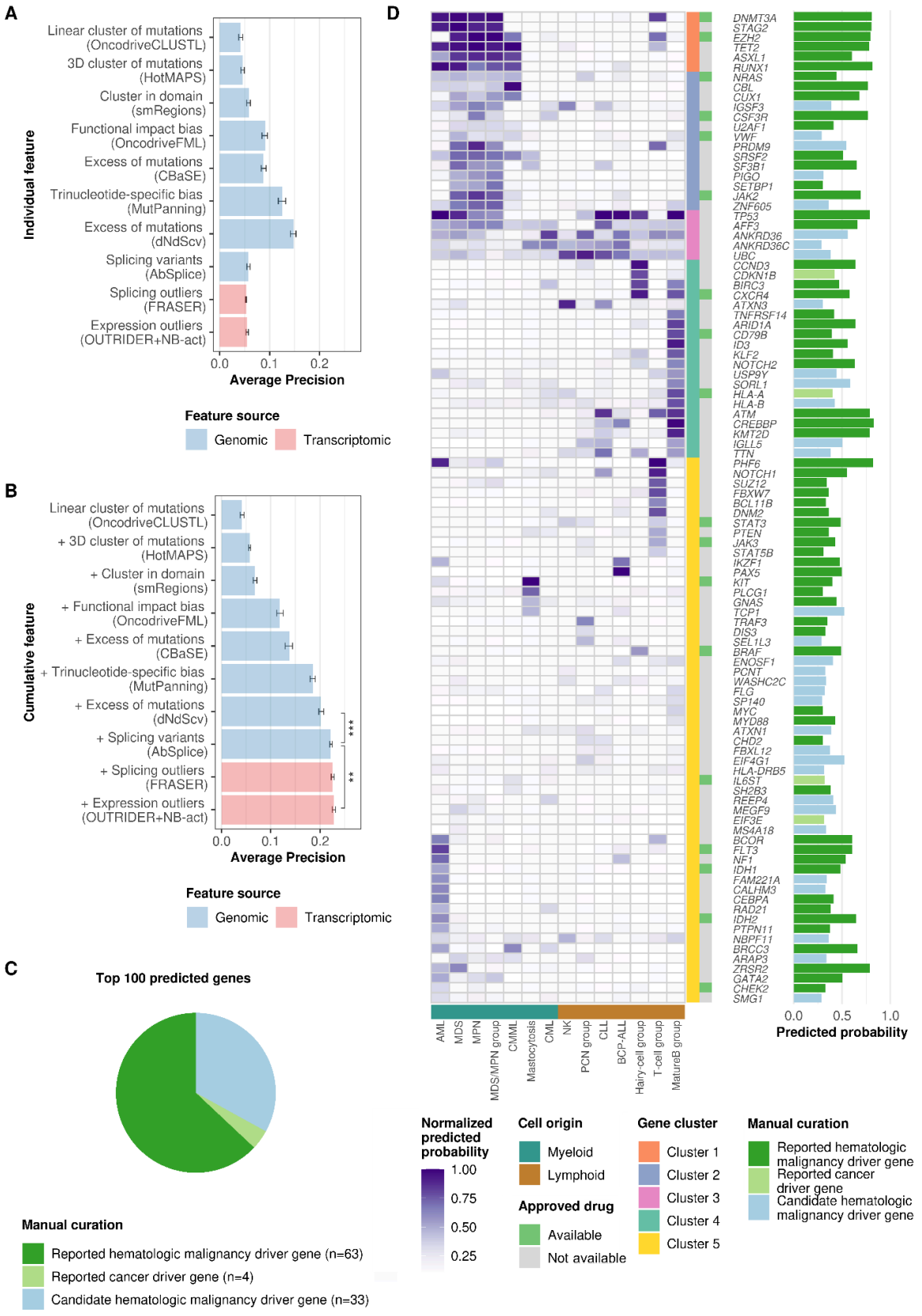


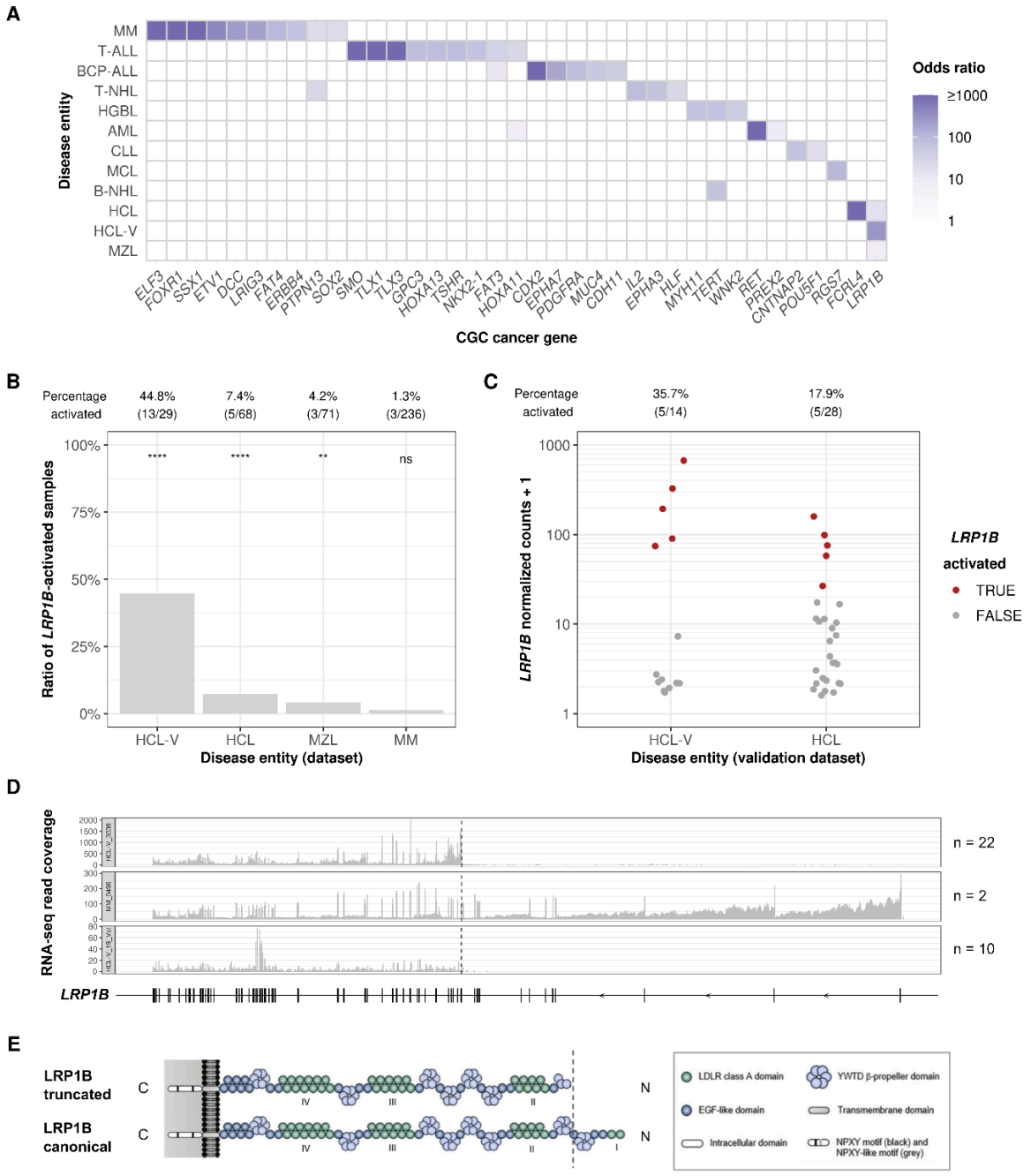
Abb. 9. Seltene abnormale genomische und transkriptomische Merkmale verbessern die Vorhersagekraft von Modellen für Krebsgene in hämatologische Malignität über die Vorhersagekraft von Methoden, die nur Muster von wiederkehrenden Mutationen analysieren. (A) Einzelne Modellmerkmale deren entsprechende Modellperformance (gemessen an der durchschnittlichen Genauigkeit) unter Verwendung eines

Random-Forest-Klassifikators. (B) Reihenfolge wie in (A) für Modellmerkmale in kumulativer Form. Die Leistung verbesserte sich konstant, wenn zusätzliche Merkmale hinzugefügt wurden. Die Sternchen kennzeichnen die nominale Signifikanz des Wilcoxon-Tests. (C) Anzahl der Gene in jeder Kategorie unter den besten 100 vorhergesagten Genen bei Verwendung aller Merkmale und des gesamten Datensatzes. Identifizierte Krebsgene hämatologischer Krankheiten sind die Gene, die entweder in CGC als Blutkrebsgene oder aus Gene-Panels für hämatologische Erkrankungen bekannt sind. Im Vergleich dazu sind Krebsgene alle von aus CGC bekannten Gene. Die übrigen Gene werden als mögliche Kandidaten kategorisiert. (D) Die Heatmap zeigt die vorhergesagte Treibergen-Wahrscheinlichkeit pro Gen (Zeilen) und Studiengruppe (Spalten) relativ zum spaltenweisen Maximalwert. Myeloide und lymphoide Entitäten wurden aufgrund der gemeinsamen hämatologischen Malignitäts-Treibergenprofile geclustert (untere Spur). Das Balkendiagramm zeigt die vorhergesagten Wahrscheinlichkeiten des auf dem gesamten Datensatz trainierten Modells. Balkenfarben wie in (C).

Einige Zusammenhänge wurden jedoch noch nicht berichtet, darunter die Überexpression von LRP1B bei Haarzellleukämie (HCL), Haarzellleukämie-Variante (HCL-V) und Marginalzonen-Lymphom (MZL). LRP1B, das Low-Density-Lipoprotein-Rezeptor-verwandte Protein 1B, ist ein häufig mutiertes Gen bei verschiedenen Krebsarten, aber seine genaue Rolle ist unklar (86). Insgesamt fanden wir in 24 Proben (0,6 % aller 3.760 Proben) eine abnormal hohe Expression von LRP1B. Davon wurden 21 innerhalb von HCL-V, HCL und MZL gefunden, wobei die LRP1B-aktivierten Proben 44,8 % (HCL-V), 7,4 % (HCL) und 4,2 % (MZL) ausmachten (Abb. 10B). Die anderen drei Fälle mit hoher LRP1B-Expression wurden in Proben mit Multiplen Myeloms (MM) gefunden. Von allen LRP1B-aktivierten Proben waren mehr als die Hälfte der Fälle (13 von 24) bei HCL-V-Patienten zu finden, was vermuten lässt, dass eine abnormale LRP1B-Expression eine wichtige Rolle spielen könnte.

Wir haben dann mit drei anderen Forschern (Christopher Oakes, Piers Blombery und Marc Seifert, siehe Abschnitt I.5) zusammengearbeitet, um diese Beobachtungen in einem unabhängigen Validierungsdatensatz von 42 Proben zu wiederholen. In Übereinstimmung mit den Beobachtungen in unserem Primärdatensatz konnten wir 10 LRP1B-aktivierte Proben in HCL-V (5/14; 36 %) und HCL (5/28; 18 %, Abb. 10C) im Validierungsdatensatz nachweisen. Darüber hinaus zeigten die RNA-seq Daten für beide Datensätze, dass Proben, die LRP1B überexprimieren, in der Mehrheit (32/34; 94 %) der Proben ein verkürztes Transkript exprimieren (22/24 im Datensatz; 10/10 im Validierungsdatensatz; Abb. 10D). Bei den beiden Proben, die Transkripte in voller Länge exprimierten, handelte es sich jeweils um ein multiples Myelom, dessen Krankheitsursachen noch ungeklärt sind. In den Fällen mit verkürzten Transkript begann die LRP1B-Expression am Exon 13, so dass die ersten 636 Aminosäuren von Exon 1 bis 12 fehlten. Drei Startcodons innerhalb von Exon 13 könnten den Beginn der Translation unter Verwendung des kanonischen offenen Leserasters ermöglichen. Angenommen, dieses Transkript wird translatiert, ergibt es ein verkürztes LRP1B-Protein, das in der Mitte der zweiten β-Propeller-Domäne beginnt (Abb. 10E). Wir konnten jedoch keine genomische Ursache für die verkürzte Isoform ausmachen, da keine einzelnen Nukleotidvarianten, kurze Insertions, kurze Deletions, strukturelle Varianten oder Genfusionen, die LRP1B betreffen, für die betroffenen Proben spezifisch waren.

Abb. 10 (Nächste Seite)**.** ***Die abnormale Aktivierung von LRP1B ist bei HCL-V vorherrschend*****.** (A) Krankheitsbilder gegen abnormal aktivierte CGC-Gene, eingefärbt gemäß ihrer Odds Ratio aus Fisher-Tests. (B) Prozentualer Anteil der LRP1B-aktivierten Proben in den vier Krankheitsbildern, in denen eine LRP1B-Aktivierung auftrat. Die Prozentsätze der Proben mit LPR1B-Aktivierung (NB-act) und die nominale Signifikanz aus dem einseitigen Fisher-Test sind oben gekennzeichnet. (C) Normalisierte Anzahl der LRP1B-aktivierten Proben des Datensatzes relativ der verschiedenen Krankheitsubtypen im Validierungsdatensatz. Die Prozentsätze der Proben, die eine LPR1B-Aktivierung aufweisen (gemäß clustering), sind oben gekennzeichnet. (D) Transkriptomische Expressionskurve, die die LRP1B-verkürzten Transkripte in den Proben HCL-V_3036 und MM_0496 (Datensatz) und HCL-V_19_Val (Validierungsdatensatz) zeigt. (E) Antizipierte Domänen anordnung von verkürztem und kanonischem LRP1B.



SP3 Krebsdiagnose und Subtypidentifizierung (Stephan Hutter)

Dieses Teilprojekt wurde unter der Leitung von Stephan Hutter in enger Zusammenarbeit mit dem Labor von Matthias Heinig bearbeitet. Die Methoden des maschinellen Lernens zur Identifizierung und Analyse der Subtypen, sowie deren Zusammenhang mit der Überlebenszeit der Patienten, wurden von Personal implementiert, das aus Teilprojekt 1 finanziert wurde.

AP3.1: Datenauswahl, Vorverarbeitung und Bereitstellung für SP1 und SP2

In der initialen Phase des Projekts wurde der Datenbestand des 5.000 Genomeprojekts dahingehend aufarbeiten, dass er den Projektpartnern zur Verfügung gestellt werden kann.

Die Daten wurden in Form von gemappte BAM-Dateien, Varianten-Calls im VCF Format und Genexpressionsmatrizen bereit gestellt.

AP3.2: Subtypidentifizierung und -diagnose

Überwachte Klassifizierung der WHO2022-Subtypen

Ziel der Analyse war es jeweils einen (multi-class) Klassifikator für die Erkennung der von der WHO2022 definierten Subtypen von AML () und MDS () zu trainieren. Für diese Analyse wurden genom- und transkriptomweite Daten für die AML- und MDS-Kohorte verwendet. Die genomischen Daten wurden binär kodiert (Vorhandensein oder Fehlen) für folgende Arten von Aberrationen: Einzelnukleotid-Varianten (AML-Kohorte n=9.947, MDS-Kohorte n=10.047), aggregiert auf Genebene und gefiltert nach mittlerem und hohem Varianteneffekt-Prädiktor (VEP), strukturelle Varianten (AML n=879, MDS n=872), und Copynumber-Varianten (AML n=842, MDS n=839). Zu den Transkriptomdaten gehören die Genexpression (AML n=25.656, MDS n=25.641) und RNA-Fusionsprodukte (AML n=7, MDS n=1).

Ein Random-Forest-Klassifikator wurde mit der scikit-learn-Bibliothek und Python erstellt. Das Modell wurde mittels fünffacher Kreuzvalidierung trainiert und evaluiert. Die Auswahl der Merkmale wurde auf Grundlage der Gini-Verunreinigung durchgeführt. Zusätzlich wurde der Beitrag einzelner Merkmale zur Klassifizierung mit Hilfe von Shapley-Werten bewertet.

Bei der Klassifizierung der acht AML-Subtypen wurde eine Genauigkeit von 0,95 erreicht. Unter 37.331 Variablen in den Eingabedaten konnten wir die 40 wichtigsten Variablen identifizieren, die für die Identifizierung des AML WHO2022-Subtyps erforderlich sind. Die Auswertung der Shapley-Werte zeigt, dass das Vorhandensein von NPM1- und CEBPA-Mutationen zur Klassifizierung von AML-NPM1 und AML-CEBPA entscheidend ist. Wie erwartet, ist eine hohe MECOM-Expression entscheidend für den AML-MECOM Subtyp, während die Strukturvariante t(8;21)(q21.3;q22.12) für AML-RUNX1::RUNX1T1 und t(15;17)(q24.1;q21.2) für den APL Subtyp wichtig ist. Der AML-CBFB::MYH11 Subtyp unterscheidet sich sowohl durch eine höhere MYH11-Genexpression als auch durch die Strukturvariante inv(16)(p13.11;q22.1). Die Klassifizierung des Subtyps AML-KMT2A beruht auf einer Kombination von veränderten Expressionsprofilen, die die Rolle von KMT2A als Regulator der Genexpression in der Hämatopoese widerspiegeln. Fehlen die oben genannten Merkmale, wird der Patient als AML-MR Subtyp klassifiziert.

In der MDS-Kohorte trainiert, erreicht das Modell eine Genauigkeit von 0,78 für die WHO2022-Subtypen, wobei 50 notwendige Merkmale beibehalten werden. Die geringere Genauigkeit ist auf Fehlklassifikationen bei den morphologisch definierten MDS-Subtypen, wie MDS-LB, MDS-IB1 und MDS-IB2, zurückzuführen. Wie bei der AML stellen wir auch in diesem Fall fest, dass das Modell bekannte Varianten wie SF3B1- und TP53-Mutationen zur Definition der jeweiligen Subtypen korrekt auswählt. Der Subtyp MDS-5q-Deletion wird durch eine negative Copynumber-variante auf dem q-Arm von Chromosom 5 korrekt identifiziert.

Diese Ergebnisse zeigen, dass genomische und transkriptomische Profile allein für eine hochpräzise AML- und MDS-Diagnose auf Subtyp-Ebene gemäß dem WHO2022-Klassifikationssystem ausreichen. Ohne vorherige manuelle Merkmalsauswahl identifiziert unser Modell bekannte diagnostische Marker korrekt als die wichtigsten Merkmale für die Klassifizierung, wie z. B. Mutationen mit mittlerem und hohem Effekt in NPM1, CEBPA, TP53 und SF3B1. Zusätzlich tragen Struktur- und Kopienzahlvarianten zur Klassifizierung bei, wie im Fall von t(8;21)(q21.3;q22.12), inv(16)(p13.11;q22.1) und 5q-assoziierten CNVs.

Unsupervised Clustering

Zur Erkennung von Merkmalen, die sich auf das Überleben der Patienten auswirken, führten wir ein unüberwachtes Clustering des gleichen Datensatzes durch, der auch für die Klassifikation der Subtypen verwendet wurde.
Das Verhältnis der Stichprobengröße zur Anzahl der Variablen beträgt sowohl bei der AML- als auch bei der MDS-Kohorten ungefähr 1:500. Dies stellt eine besondere Schwierigkeit für unüberwachte Methoden wie K-Means oder Gaussian Mixture Models (GMM) dar, da diese Methoden aufgrund einer großen Anzahl irrelevanter Variablen instabile Cluster erzeugen können. Um die Anzahl der Variablen zu reduzieren, nutzten wir mehrere Strategien, einschließlich der Auswahl der Variablen mit höchster Varianz und der wichtigsten Variablen für die Klassifikation von AML oder MDS Patienten gegen den Rest des Datensatzes. Die zweite Strategie führte zu einer Auswahl von SNP-Mutationen und Strukturvarianten, die auch routinemäßig in der Diagnostik der WHO2022-Subtypen verwendet werden, sowie zu einer Untergruppe von Genexpressionsmerkmalen.
Auf der Grundlage von N=20 mittels der zweiten Strategie ausgewählten Variablen wurden drei Clustering-Algorithmen angewendet: K-means, GMM und ein generalisiertes Mischmodell. Die sich ergebenden Cluster wurden intern anhand der Silhouette-width und des Bayes'schen Informationskriteriums (BIC) und extern durch Berechnung der Genauigkeit der WHO2022-Subtypen bewertet.
Die besten Clustering-Ergebnisse wurden mit K-means und GMM erzielt. Bei der AML-Kohorte (n=621) erzeugte der GMM-Algorithmus 7 verschiedene Cluster, die mit einer Genauigkeit von 76 % mit den WHO2022-Subtypen übereinstimmten (Abb. 11). Die Cluster sind durch AML-Mutationen charakterisiert, die für den entsprechenden WHO2022-Subtyp typisch sind. Dazu zählen u.a. NPM1 und CEPBA Mutationen, KMT2A-/ und MECOM-rearrangements, sowie spezifische Genfusionen. Die Patienten ohne AML-definierende genetische Veränderungen sind als AML-MR klassifiziert. Vier der 7 Cluster zeigten eine homogene Zusammensetzung: c2 bestand ausschließlich aus AML mit KMT2A-rearrangements, c3 ausschließlich aus AML mit MECOM-rearrangements, c5 ausschließlich aus AML mit NPM1 Mutationen, und c6 ausschließlich aus AML mit RUNX1::RUNX1T1 Fusionen. Die übrigen 3 Cluster beinhalteten mehrere WHO2022-Subtypen.
In der MDS-Kohorte (n=698) führte kein Ansatz zu einer klaren Clusterstruktur. Bei ausschließlicher Anwendung auf die durch Blasten definierten Subtypen (MDS-LB, MDS-IB1, MDS-IB2) identifizierte der K-Means-Ansatz jedoch drei verschiedene Cluster (Silhouettenbreite von 0,98), von denen zwei durch Mutationen von STAG2 (n=28) und U2AF1 (n=34) definiert wurden, während die Mehrheit der Patienten keine besondere Signatur aufweist (n=346, Cluster 3).
Zusammenfassend lässt sich feststellen, dass die unüberwachte datengesteuerte Analyse zu Clustern führt, die der WHO2022 AML-Klassifikation sehr nahe kommen. Dies ebnet den Weg zur Definition von AML-Subtypen allein auf der Grundlage von Sequenzierungsmethoden.

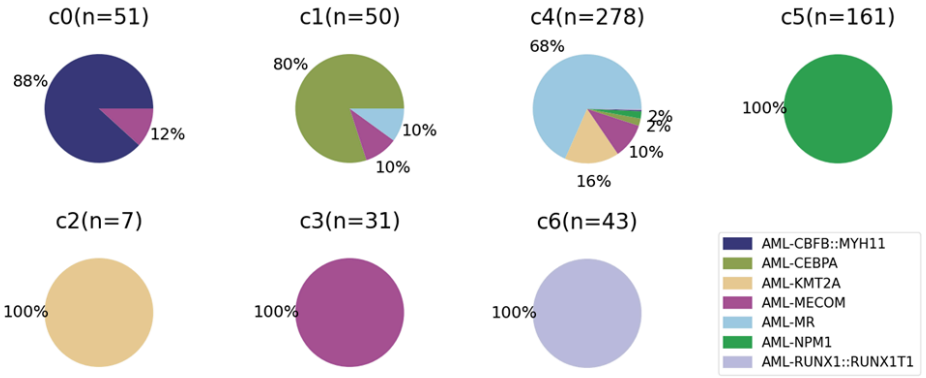


Abb. 11. *Clustering der AML-Kohorte***.** Mittels GMM-Algorithmus wurden 7 verschiedene Cluster erzeugt. Farben spiegeln die jeweiligen WHO2022-Subtypen wieder.

AP3.3: Integrierung molekularer Profile in die Risikostratifizierung des Patienten und die Vorhersage des Behandlungsergebnisses

Auf Grundlage der Ergebnisse des Clustering in 3.2 wurde untersucht, ob diese unterschiedlichen molekularen Profile der Patienten dazu geeignet sind, eine Prognose über deren Mortalitätsrisiko zu treffen. Dazu verwendeten wir ein Cox-Regressionsmodel um die Zeit bis zum Todesfall in Abhängigkeit der Zugehörigkeit zu einem der Cluster zu modellieren.

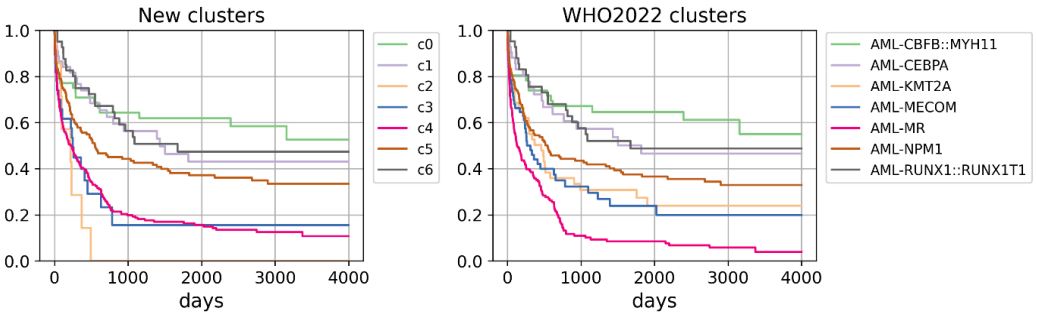


Abb. 12. Überlebensanalyse der AML-Kohorte. Überlebensanalyse der AML-Kohorte stratifiziert nach neuen Cluster (links) und nach WHO2022-Subtypen (rechts).

In der AML Kohorte ergab die Überlebensanalyse einen C-Index von 0,6 (Abb. 12, links). Diese Ergebnis ist dem Ergebnis der Überlebensanalyse auf Grundlage der WHO2022-Subtypen sehr ähnlich (C-Index von 0,62, Abb. 12, rechts). Wie Abb. 10 zu entnehmen ist, entsprechen die neuen Cluster großteils der WHO2022-Subtypen. Interessanterweise, zeigt das Cluster c2, welches KMT2A-rearrangierten AMLs enthält, ein deutlich schlechteres Überleben. Dies deutet darauf hin, dass das Clustering in der Lage ist, eine Untergruppe der AML mit KMT2A-rearrangements zu identifizieren, die eine deutlich schlechtere Prognose zeigen bezüglich Gesamtüberleben, als die übrigen KMT2A-rearrangeierten AMLs. Es sind jedoch weitere Analysen notwendig, um diese Fälle genauer zu charakterisieren.

In der MDS Kohorte führte die Überlebensanalyse der neuen Cluster zu einem niedrigeren C-Index (0,54) im Vergleich zur WHO2022-basierten Vorhersage (C-Index von 0,59). Dies liegt vermutlich daran, dass es bei MDS im Vergleich zu AML wesentlich weniger klar genetisch-definierte Subgruppen gibt und häufige genetische Aberrationen oft in den unterschiedlichsten MDS Subgruppen zu finden sind. Somit stellt die genetische Charakterisierung des MDS eine zusätzliche Herausforderung dar.

Insgesamt lässt sich sagen, dass die Überlebensanalyse keine sehr genauen Vorhersagen zulässt. Dies ist in beiden Kohorten der Fall und gilt sowohl für die Subtypen, die durch das Clustering definiert wurden, als auch für die Subtypen, die durch die WHO definiert wurden.
Kürzlich wurde ein MDS Risiko-Stratifizierungsmodell etabliert (IPSS-M), welches unabhängig von MDS Subtypen molekulare Daten in das Modell integriert (Bernard et al. Molecular International Prognostic Scoring System for Myelodysplastic Syndromes. NEJM Evidence. 2022; 1(7)). Hier zeigte sich, dass solche Modelle im MDS eher das Leukämiefreie-Überleben vorhersagen können, anstatt das Gesamtüberleben. In einem Folgeprojekt könnte daher die hier beschriebenen ML Methoden im Kontext dieses Endpunkts angewandt werden.

2. Wichtigste Positionen des zahlenmäßigen Nachweises

Die Förderung umfasste 357084,25 Euro. Davon entfielen 247250,05 Euro auf Personalkosten (Postdoc mit Schwerpunkt Teilprojekt 1). Die Reiskosten betrugen 468,90 Euro. Für das Projekt wurden 75579,65 Euro an Sachmitteln für projektspezifische Rechner-Infrastruktur ausgegeben. Die Verwaltungskosten betrugen 33785,65 Euro.
Die Mittel wurden wie geplant eingesetzt.

3. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Alle Arbeiten waren wesentlich, notwendig und hinreichend für den Erfolg des VALE Projekts.

4. Voraussichtlicher Nutzen, insbesondere Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans mit Zeithorizont

Wirtschaftliche Erfolgsaussichten		Während der Laufzeit des Vorhabens	Im Anschluss an die Laufzeit
Ziel	**Erläuterungen**		
Volkswirtschaftliche Verwertung	Mögliche Einsparungen im Gesundheitswesen durch akurate Molekulardiagnostik (Molecular Tumour boards)	-	x

Wissenschaftliche und/oder technische Erfolgsaussichten		Während der Laufzeit des Vorhabens	Im Anschluss an die Laufzeit
Ziel	**Erläuterungen**		
Verbreitung der Erkenntnisse	Veröffentlichung der erzielten Ergebnisse in wissenschaftlichen Fachzeitschriften und für die breite Öffentlichkeit, Meldung bei Studienregistern, Vorträge und Poster bei Fachkongressen oder anderen Veranstaltungen	x	x

Aus-, Weiter-, Fortbildung	Vorhaben dient der Heranbildung des wissenschaftlichen Nachwuchses, durch die Erstellung von Dissertationen, die Ergebnisse finden Eingang in die Lehre	x	x
Forschungsstrukturen	Entwicklung und Instandhaltung von robuster Software und einer Pipeline für die Detektion von aberranten Transkriptions-Ereignissen.	x	x

Wissenschaftliche und wirtschaftliche Anschlussfähigkeit		Im Anschluss an das Vorhaben
Ziel	**Erläuterungen**	*Zeithorizont erläutern*
Wissenschaftliche, technische, strukturelle und versorgungsbezogene Verwertungsmöglichkeiten	Die Methodenentwicklung auf dem Gebiet der aberranten Transkriptom-Ereignissen in Krebsproben wird über das Projektende hinaus notwendig sein. Die in diesem Projekt erzielten Fortschritte bilden die Basis für weitere Forschungsprojekte und Anträge: GHGA (https://www.ghga.de/) zweite Förderperiode, grant, ERDERA (https://www.ejprarediseases.org/erdera/); BMBF “Nationalen Dekade gegen Krebs” Datenanalyse und des Datenteilens in der Krebsforschung; EXIST Gründungsstipendium	ERDERA: 09.2024 EXIST Gründungsstipendium: bewilligt, voraussichtlicher Start 09.2024 BMBF - Dekaden gegen Krebs: Antrag überreicht GHGA: 10.2025

5. Während der Durchführung des Vorhabens bekannt gewordenen Fortschritts auf dem Gebiet des Vorhabens bei anderen Stellen

Es sind von dritter Seite keine Ergebnisse bekannt geworden, die wesentlichen Einfluss auf die Verwertung der Ergebnisse nehmen.

6. Erfolgte oder geplante Veröffentlichungen des Ergebnisses

Teilprojekt 1 unter Federführung von Matthias Heinig:

Vilov S, Heinig M. DeepSom: a CNN-based approach to somatic variant calling in WGS samples without a matched normal. Bioinformatics. 2023; 39(1):btac828. doi: 10.1093/bioinformatics/btac828.

Vilov S, Heinig M. Investigating the performance of foundation models on human 3’UTR sequences. bioRxiv 2024; doi: 10.1101/2024.02.09.579631

Teilprojekt 2 unter Federführung von Julien Gagneur:

Bredthauer C, Fischer A, Ahari AJ, Cao X, Weber J, Rad L, Rad R, Wachutka L, Gagneur J. Transmicron: accurate prediction of insertion probabilities improves detection of cancer driver genes from transposon mutagenesis screens. Nucleic Acids Research, 2023

Hölzlwimmer FR, Lindner J, Wagner N, Yépez VA, Casale FP, Gagneur J. Aberrant expression prediction across human tissues. bioRxiv, 2023

Teilprojekt 1-3 unter Federführung von Julien Gagneur und Stephan Hutter:

Cao X, Huber S, Ahari AJ, Traube FR, Seifert M, Oakes CC, Secheyko P, Vilov S, Scheller IF, Wagner N, Yépez VA, Blombery P, Haferlach T, Heinig M, Wachutka L, Hutter S, Gagneur J. Analysis of 3760 hematologic malignancies reveals rare transcriptomic aberrations of driver genes. Genome Med. 2024; 16(1):70. doi: 10.1186/s13073-024-01331-6.

München, den ……..... …….....

Dr. Matthias Heinig