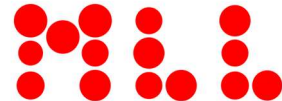


GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

MLL Münchner Leukämielabor GmbH
Max-Lebsche-Platz 31
81377 München



**Schlussbericht zum CompLS – Runde 2 – Verbundprojekt:
VALE – Entdeckung und Vorhersage der Wirkung von
genetischen Varianten durch Artifizielle Intelligenz für
LEukämie Diagnose und Subtyp-Identifizierung –
Teilproject C**

Verantwortlich: Dr. rer. nat. Stephan Hutter

Förderkennzeichen: 031L02003C

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

I. Kurzbericht

1. Aufgabenstellung, sowie wissenschaftlicher und technischer Stand, an den angeknüpft wurde

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 **e**ffect 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, vom Münchner Leukämielabor (MLL), betreut eine einzigartige Genomressource, die fast 5.000 vollständige Genome (WGS, Whole Genome Sequencing) und entsprechende Transkriptome (RNA-Seq) von Leukämiepatienten sowie weitere Informationen zu Behandlungsergebnissen und Überleben umfasst (MLL5K Datensatz).

Der MLL5K Datensatz wurde erzeugt, um das Verständnis der molekularen Grundlagen von Leukämien und Lymphomen zu verbessern. Er umfasst ein breites Spektrum hämatologischer Malignome, von häufigen Entitäten wie akuter myeloischer Leukämie (AML) oder myelodysplastischem Syndrom (MDS) bis hin zu seltenen Krankheiten wie der 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, und der MLL5K Datensatz wurde seitdem von vielen weiteren Forschungsgruppen genutzt um eine Vielzahl an diversen hämatologischen Fragestellungen zu bearbeiten. Im Rahmen des VALE Verbunds wurden von den Arbeitsgruppen Heinig und Gagneur entwickelte Methoden und Modelle aus dem maschinellen Lernen (ML) auf den MLL5K Datensatz angewandt und deren Ergebnisse von Wissenschaftlerinnen des MLL aus hämatologischer Sicht analysiert und interpretiert.

2. Ablauf des Vorhabens

Die kombinierte Analyse 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. Anschließend soll analysiert werden,

inwiefern die molekularen Profile der Patienten neue Subgruppen definieren und wie sich diese in Bezug auf Überleben oder Behandlungsergebnis der Patienten verhalten.

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

SP1 Varianten-Identifikation und Varianten-Effektvorhersage (Leitung 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 (Leitung Julien Gagneur)

- 2.1 Expressions- und Spleißausreißer im MLL5K RNA-Seq Datensatz
- 2.2 Priorisierung von Cis-regulatorischen Varianten
- 2.3 Priorisierung von Kandidatengenen für die Metaanalyse über die Proben hinweg

SP3 Krebsdiagnose und Subtypidentifizierung (Leitung 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

Die Bearbeitung aller dieser Fragestellungen erfolgte in enger Zusammenarbeit zwischen allen drei Arbeitsgruppen, wobei die wesentlichsten Beiträge unseres Teilvorhabens sich auf Teilprojekte SP2 und SP3 konzentrierten.

3. Wesentliche Ergebnisse, sowie Zusammenarbeit mit anderen Stellen

Durch die gemeinsame Analyse von Expressions- und Spleißausreißer auf RNA-Ebene, sowie von genomischen Varianten auf DNA-Ebene (SP2) konnten Gene identifiziert werden, die mit bestimmten Leukämietypen assoziiert sind. Dies beinhaltete neben bereits etablierten Treiber-Genen auch Gene, die bis jetzt noch nicht als relevant für die Krankheitsentwicklung beschrieben wurden. Diese neuen Gene repräsentieren interessante Kandidaten für weitere Untersuchungen. Ein besonders vielversprechendes Gen hierbei ist *LRP1B*. Dieses Gen ist spezifisch für eine Untergruppe von seltenen B-Zellerkrankungen über die auf molekulargenetischer Ebene noch wenig bekannt ist.

In SP3 konnte gezeigt werden, dass ML Modelle, die alleine auf genomische und transkriptomische Daten angewandt wurden, mit hoher Präzision die jeweiligen Patienten in die von der WHO definierten Subtypen von AML und MDS einteilen können. Darüber hinaus konnten unüberwachte ML Methoden neue Unterklassen entdecken, die ein unterschiedliches Überleben aufweisen. So konnte z.B. eine Gruppe innerhalb der *KMT2A* definierten AML identifiziert werden, welche sich durch einen besonders schlechten Progress auszeichnet.

Zusammenarbeiten im Rahmen der Kooperation erfolgten mit 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 zu

1. der Verwendung der Zuwendung und den erzielten wissenschaftlich-techn. Ergebnissen im Einzelnen, mit Gegenüberstellung zu den ursprünglichen Zielen,

Da bei Bearbeitung sämtlicher Arbeitspakete im Rahmen des Vorhabens in enger Zusammenarbeit zwischen den einzelnen Arbeitsgruppen erfolgte, beschreiben wir im Folgenden die Ergebnisse der Arbeitspakete bei denen wesentliche Leistungen mit den Mitteln des Teilprojekts erbracht wurden.

Dies beinhaltet neben den Arbeitspaketen des Subprojektes SP3 auch die Arbeitspakete von SP2.

SP2 Driver-Identifikation aus RNA-Sequenz und WGS

In Teilprojekt SP2 wurde unter der Leitung von Julien Gagneur WGS-und RNA-seq-Daten von 3.760 Tumورproben untersucht haben, die aus insgesamt 24 verschiedenen Leukämie- und Lymphom-Typen stammen (Abb. 1). 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. Die Ergebnisse aus diesen Analysen wurden in Genome Medicine veröffentlicht (Cao et al. 2024).

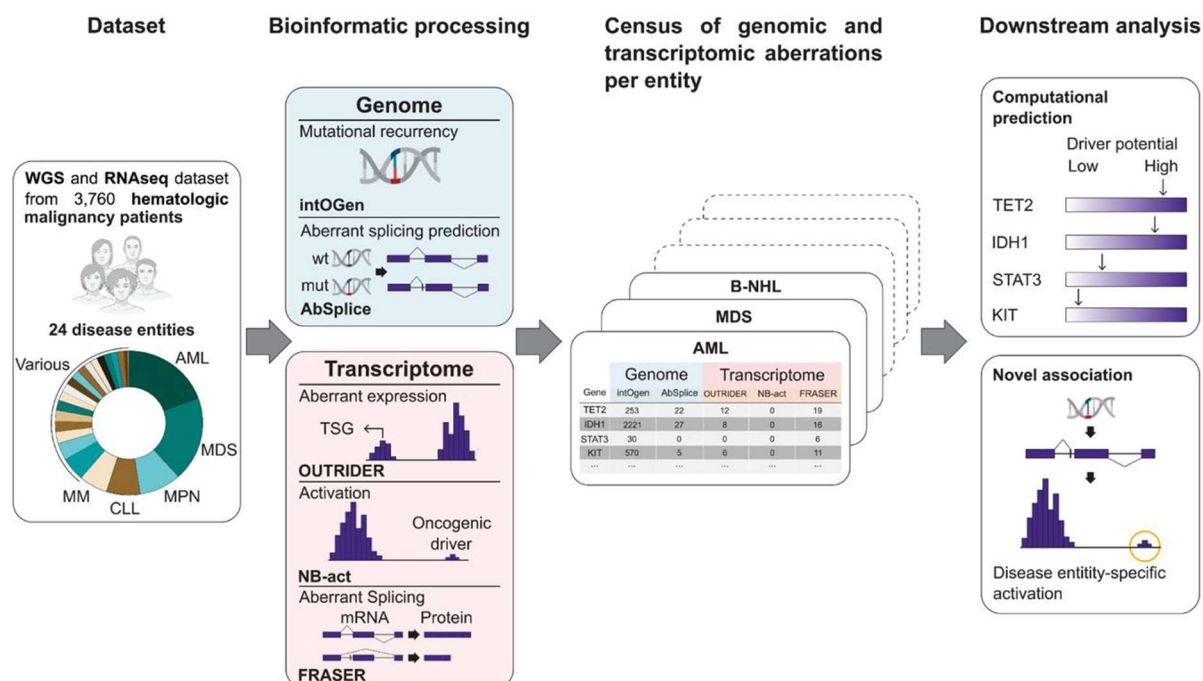


Abb 1. Schema der Hauptstudie. Datensatz: WGS und RNA-seq Daten von 3.760 hämatologischen Malignomen, die 24 verschiedene Krankheits Subtypen umfassen. Bioinformatische Verarbeitung: intOGen detektiert wiederkehrende Mutationsmuster in genomischen Daten und AbSplice prognostiziert Varianten, die abnormales Spleißen verursachen. OUTRIDER detektiert abnormale Expression häufig exprimierter Gene, NB-act detektiert die

Überexpression selten exprimierter Gene ab (für die Studie entwickelt), und FRASER identifiziert abnormales Spleißen. 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.

Wesentliche Beiträge zu dieser Arbeit wurden durch das für unser Teilprojekt bewilligte Personal erbracht. Dies beinhaltete:

- *Einteilung der Tumorproben in für die Analysen geeignete Subgruppen:* Da ein Großteil der Interpretation der Daten im Kontext von biologisch unterschiedlichen Gruppen erfolgte, wurden sämtliche Proben auf ihre Diagnosen hin überprüft und in geeignete hämatologische Subgruppen unterteilt. Dabei war es wichtig die Gruppengröße groß genug zu halten um eine statistisch sinnvolle Analyse zu ermöglichen wobei die Gruppen gleichzeitig möglichst homogen bezüglich ihrer Tumorbilogie bleiben sollten.
- *Biologische Interpretation der Ergebnisse:* Ergebnisse der unterschiedlichen Datenanalysen wurden zunächst auf ihre biologische Sinnhaftigkeit hin überprüft. Dies half beim Auffinden von Bugs und bei der generellen Optimierung der Analyse-Pipelines. Dies beinhaltete auch das Etablieren von Filtereinstellungen um technische Artefakte aus dem MLL5K Datensatz zu entfernen. Die finalen Ergebnisse wurden anschließend in einem hämatologisch/onkologischen Kontext interpretiert.
- *Literaturrecherche zur Relevanz der Ergebnisse:* Viele Gene, die als Kandidaten im Rahmen der Analysen gefunden wurden, waren bisher nicht mit hämatologischen Erkrankungen assoziiert. Es wurde daher eine umfangreiche Literaturrecherche zu diesen Genen durchgeführt um ihre funktionelle Relevanz zur Tumorbilogie einschätzen zu können, sowie um ihr diagnostisches Potential zu ermitteln.

Im Folgenden werden die wesentlichen Ergebnisse der gemeinsamen Arbeiten zu den Arbeitspaketen im Rahmen von SP2 zusammengefasst:

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

Expressionsausreißer mit OUTRIDER

Die MLL-Daten wurden mit Hilfe der Software DROP (Yépez et al. 2021) 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 (<https://cancer.sanger.ac.uk/census>) übernommen haben, gemessen. Wir fanden

eine starke Häufung für Tumorsuppressorgene unter den Unterexpressionsausreißern und für Onkogene unter den Überexpressionsausreißern (Abb. 2). 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.

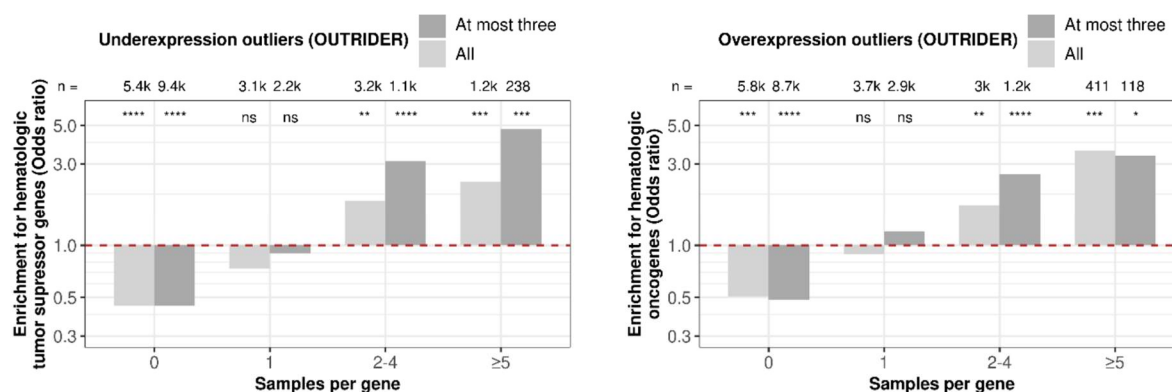


Abb. 2 Expressionsausreißer finden sich gehäuft in Onkogenen hämatologischer Erkrankungen. (Links) Häufung von aus CGC bekannten Tumorsuppressorgenen hämatologischer Erkrankungen unter allen Genen, die von OUTRIDER als unterexprimiert eingestuft wurden, sowie unter den TOP drei signifikanten Genen pro 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; *: $P \leq 0.05$; **: $P \leq 0.01$; ***: $P \leq 0.001$; ****: $P \leq 0.0001$). (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 (Abbildung 3). Auch hier erhöhte die Beschränkung auf höchstens drei Ausreißer pro Probe deren Häufigkeit.

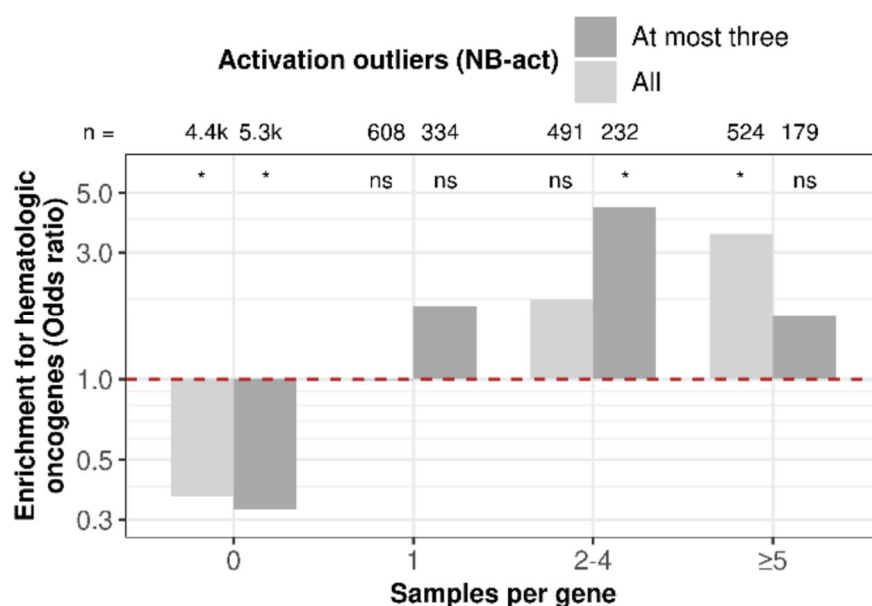


Abb. 3. Aktivierungsausreißer finden sich gehäuft in Onkogenen hämatologischer Erkrankungen. Wie in Abb. 2, jedoch für Gene, deren Aktivierung mit NB-act nachgewiesen wurde.

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 der Gruppe Gagneur wurde eine verbesserte Version von FRASER, FRASER 2.0 (Scheller et al. 2023), entwickelt, die wir nun auf den gesamten Datensatz angewendet haben.

Wir wendeten FRASER an, um abnormale 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 Tumorsuppressorgenen hämatologischer Erkrankungen unter diesen Spleißausreißern (Abb. 4). 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.

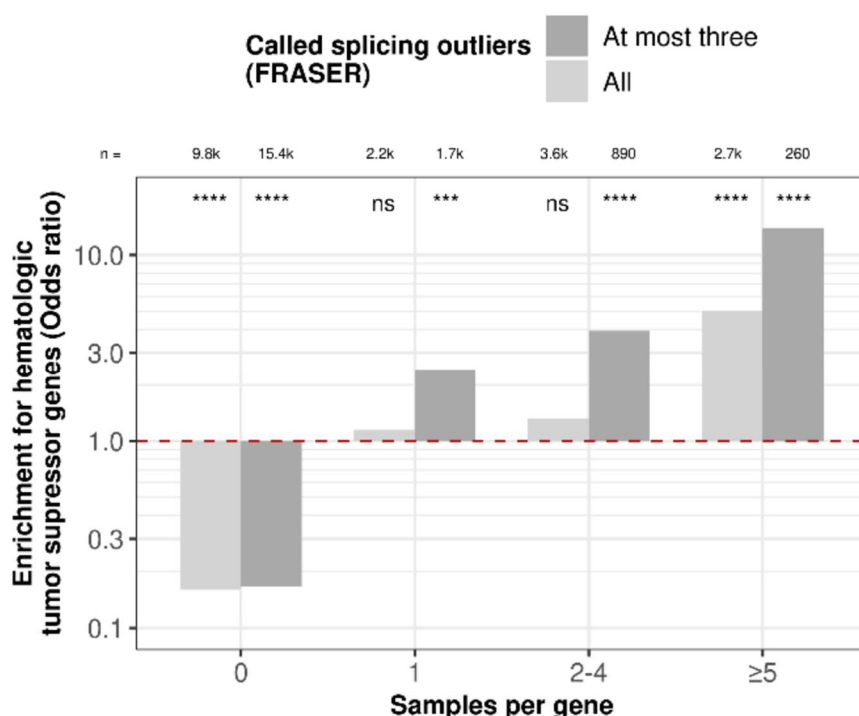


Abb. 4 Abnormales Spleißen finden sich gehäuft in Krebsgenen hämatologischer Erkrankungen. Häufung von aus CGC bekannten Tumorsuppressorgenen hämatologischer Erkrankungen unter allen von FRASER identifizierten Genen und unter den drei am höchsten signifikanten Genen pro Probe, die von FRASER als abnormal 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 et al. 2020) 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 eine Häufung von somatischen Mutationen in Tumoren sowohl in kodierenden als auch in nicht-kodierenden Genomregionen identifiziert.

In einem parallel laufenden Projekt hat die Gruppe Gagneur eine neue Methode, genannt AbSplice, zur Vorhersage von Genvarianten, welche falsches Spleißen zur Folge haben,

entwickelt (Wagner et al. 2023). 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. 5).

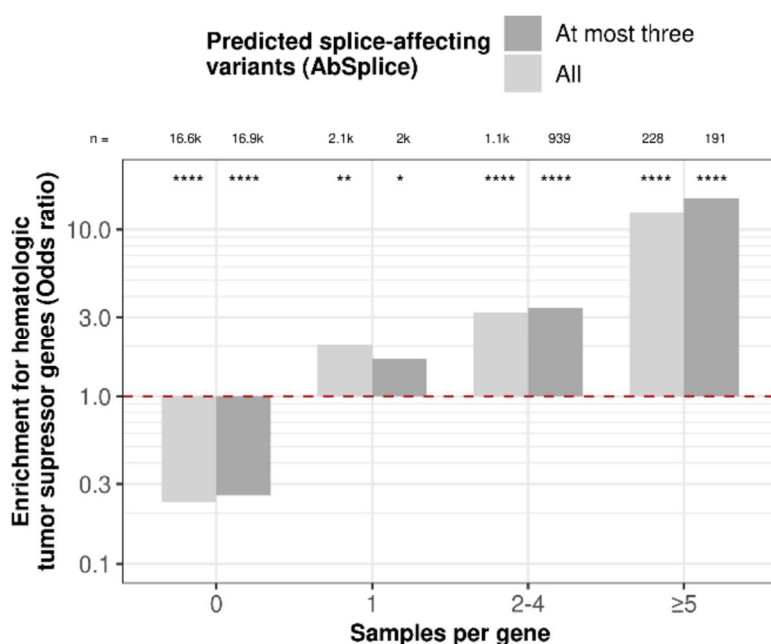


Abb. 5 Varianten, die das Spleißen beeinträchtigen, finden sich gehäuft in Krebsgenen hämatologischer Erkrankungen. Ähnlich wie in Abb. 5, aber für Varianten, die das Spleißen beeinträchtigen und mit AbSplice vorhergesagt wurden.

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

Treibergen Vorhersagen

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. 6A-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. 6B). 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, $P = 1,6 \times 10^{-84}$, Fisher-Test; Abb. 6C). 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. 6D). Ein Clustern der 100 bestplatzierten Gene nach den von den Studiengruppen vorhergesagten Wahrscheinlichkeiten ergab verschiedene krankheitsspezifische Spezialfälle (siehe Cao et al. 2024).

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 et al. 2012). *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 et al. 2022). 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.

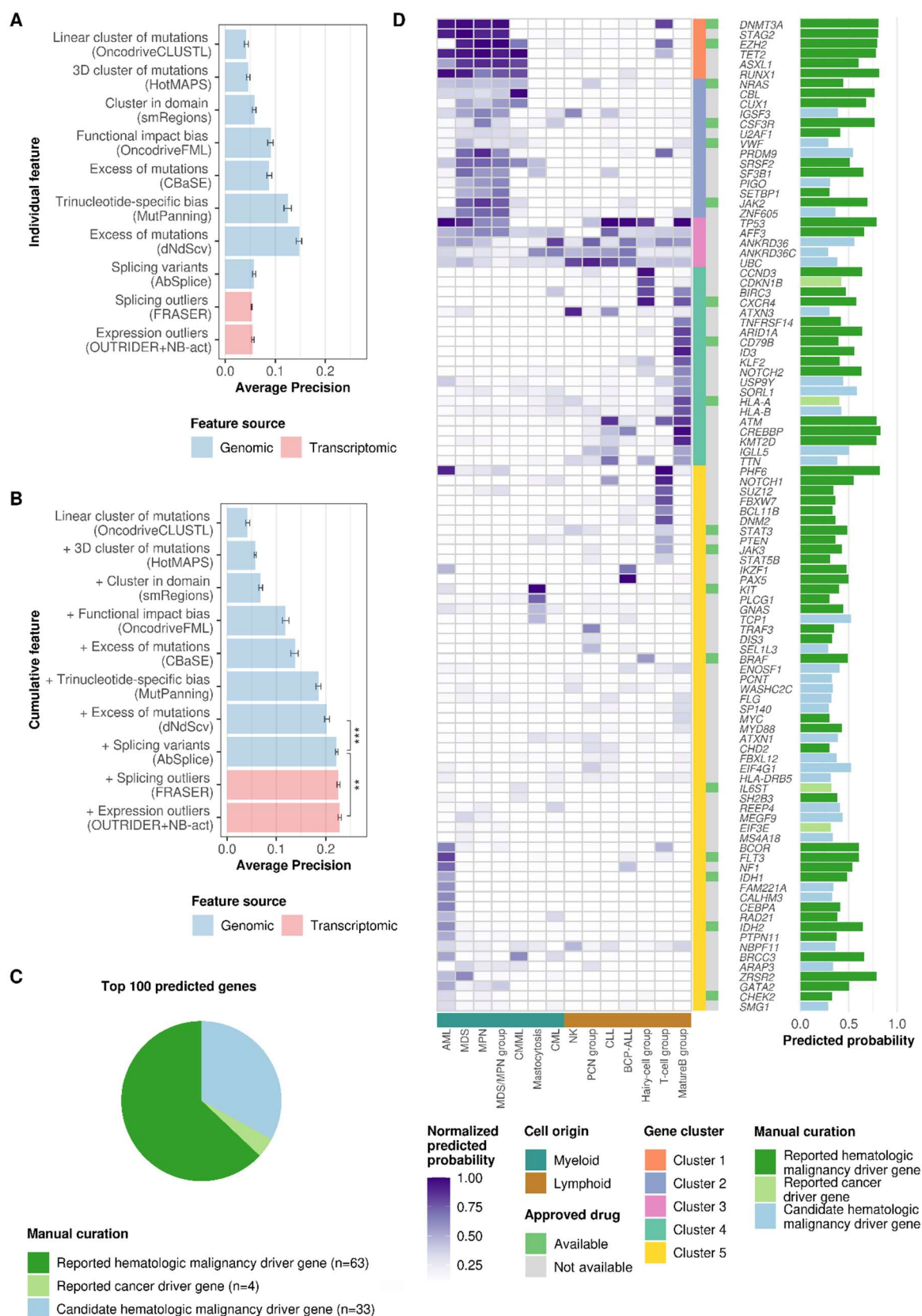


Abb. 6. Seltene abnormale genomische und transkriptomische Merkmale verbessern die Vorhersagekraft von Modellen für Krebsgene in hämatologische Malignität über die Vorhersagekraft von Methoden, die nur Muster von wiederkehrenden Mutationen analysieren. (A) Einzelne Modellmerkmale deren entsprechende

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).

***LRP1B*-Subtyp innerhalb splenic B-cell lymphoma/leukemia with prominent nucleoli (SBLPN) nach WHO2022**

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-Krebstreibergen und 12 Krankheitsentitäten (Abb. 7A). Einige Assoziationen wurden bereits in der Literatur beschrieben (Einzelheiten siehe Originalveröffentlichungen), was die Ergebnisse untermauert.

Einige Zusammenhänge wurden jedoch noch nicht berichtet, darunter die Überexpression von *LRP1B* bei Haarzelleukämie (HCL), Haarzelleukämie-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. 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. 7B). 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 %, Abbildung 7C) 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. 7D). 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. 7E). Wir konnten jedoch keine genomische Ursache für die verkürzte Isoform ausmachen, da keine einzelnen Nukleotidvarianten, kurze Insertionen, kurze Deletionen, strukturelle Varianten oder Genfusionen, die *LRP1B* betreffen, für die betroffenen Proben spezifisch waren.

Die hohe Expression von *LRP1B* beschränkte sich auf einige Fälle von Marginalzonen-Lymphomen (MZL), Haarzell-Leukämie (HCL) und der Mehrzahl der HCL-V-Fälle. Diese drei Krankheitsentitäten sind reife B-Zell-Malignome und wurden früher in der überarbeiteten 4. Auflage der WHO-Klassifikation hämatolymphoider Tumoren als separate Entitäten betrachtet, die anhand immunphänotypischer Marker und molekulargenetischer Merkmale unterschieden wurden. HCL-V ist resistent gegen die herkömmliche HCL-Therapie und weist keine HCL-spezifische BRAF-Mutation auf. Da HCL, HCL-V und MZL jedoch aus malignen reifen B-Zellen mit ähnlicher Morphologie entstehen, ist eine klare Unterscheidung oft nicht möglich. Daher wurde in der kürzlich veröffentlichten 5. Auflage der WHO-Klassifikation (WHO 2022) der Begriff "HCL-V" gestrichen, um der Tatsache Rechnung zu tragen, dass die Biologie dieser Krankheit nicht mit HCL verwandt ist. Stattdessen werden diese Fälle nun als *splenic B-cell lymphoma/leukemia with prominent nucleoli* (SBLPN) bezeichnet, die aufgrund ähnlicher zytomorphologischer Merkmale auch seltene Fälle von splenische MZL und B-prolymphozytärer Leukämie umfasst. SBLPN dient eher als Platzhalter für jene morphologisch definierten Fälle von B-Zell-Lymphomen, die sich nach derzeitigem Kenntnisstand nicht in biologisch unterschiedliche Entitäten einordnen lassen.

Unsere Daten deuten darauf hin, dass *LRP1B* möglicherweise als Marker für die Unterscheidung zwischen HCL-V und MZL innerhalb von SBLPN dienen könnte, wobei eine Überexpression von *LRP1B* auf eine schlechtere Prognose hindeutet. Weitere Untersuchungen sind jedoch erforderlich, um festzustellen, ob *LRP1B* ein treibendes Gen ist. *LRP1B*, auch bekannt als Low-Density-Lipoprotein-Rezeptor-verwandtes Protein 1B, wird in zahlreichen normalen Geweben breit exprimiert. Es spielt eine Rolle bei verschiedenen biologischen Prozessen wie Angiogenese, Chemotaxis, Proliferation, Adhäsion, Apoptose, Endozytose, Immunität, Wirt-Virus-Interaktion und Proteinfaltung. Darüber hinaus gehört *LRP1B* zu den am häufigsten veränderten Genen bei menschlichen Krebserkrankungen insgesamt. Es wird häufig durch verschiedene genetische und epigenetische Mechanismen inaktiviert, was es zu einem mutmaßlichen Tumorsuppressor-Gen macht. Diese früheren Beobachtungen stehen nicht unbedingt im Widerspruch zu unseren Ergebnissen, da wir hier eine Überexpression einer verkürzten Isoform von *LRP1B* in Zellen festgestellt haben, die es normalerweise nicht exprimieren. Trotz der Herausforderungen, die die enorme Größe von *LRP1B* mit sich bringt, wie in der Studie von Príncipe et al. hervorgehoben wurde, ermutigen unsere Ergebnisse zu weiteren Untersuchungen, um die potenzielle Rolle von *LRP1B* bei B-Zell-Malignomen zu entschlüsseln.

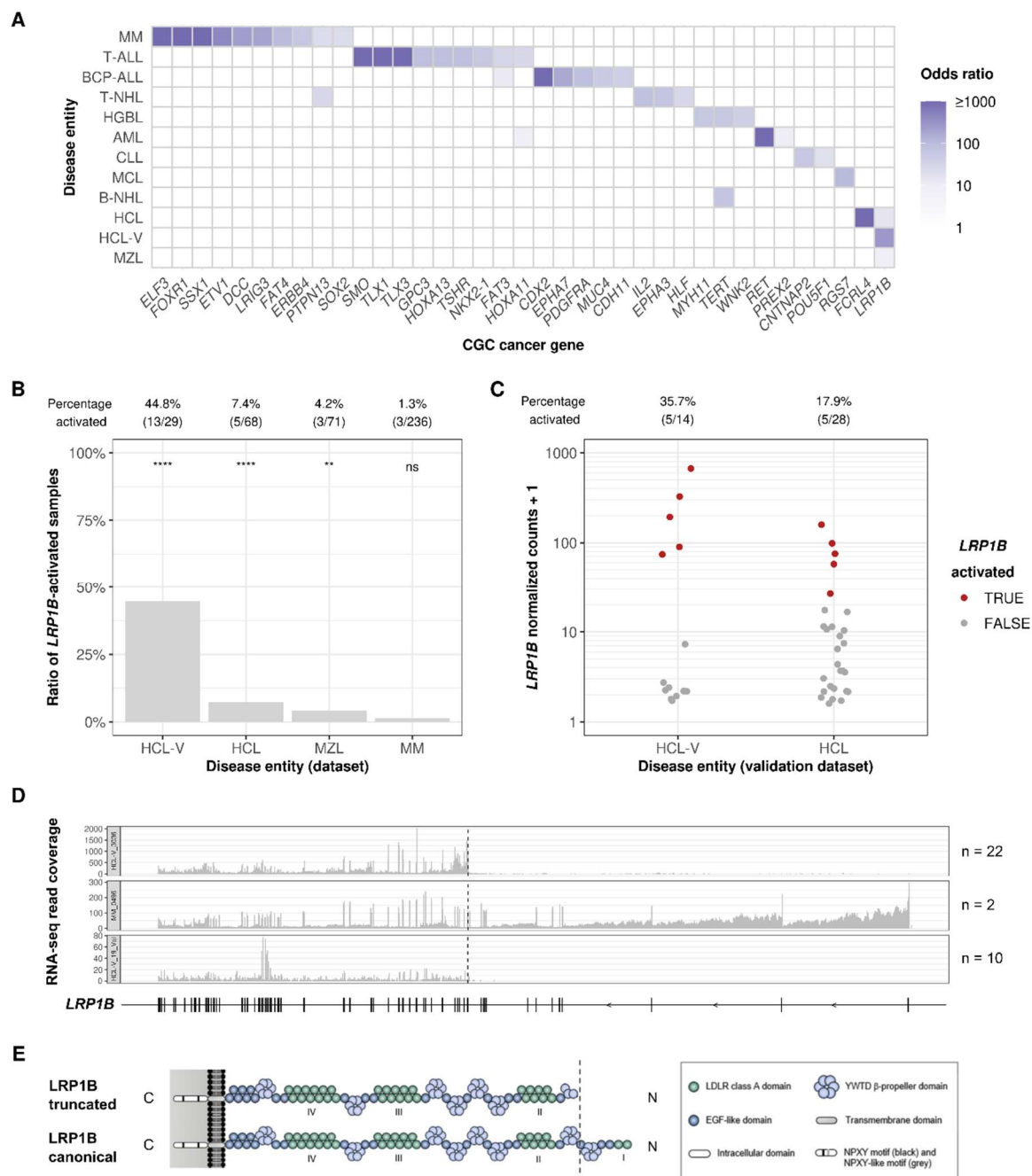


Abb. 7. Die abnormale Aktivierung von *LRP1B* ist bei HCL-V vorherrschend. (A) Krankheitsbilder gegen abnormal aktivierte CGC-Gene, eingefärbt gemäß ihrer Odds Ratio aus Fisher-Tests. (B) Prozentualer Anteil der *LRP1B*-aktivierten Proben in den vier Krankheitsbildern, in denen eine *LRP1B*-Aktivierung auftrat. Die Prozentsätze der Proben mit *LRP1B*-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 *LRP1B*-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

Die Bearbeitung des Teilprojekts SP3 fand unter der Leitung von Stephan Hutter statt. Dies beinhaltete die Datenaufbereitung des MLL5K Rohdatensatzes um diesen für die Analysen in SP1 und SP2 verfügbar zu machen (AP3.1).

Zusätzlich wurden Analysen zur Subtypidentifizierung und Überlebensvorhersage auf Basis der molekularen Daten des MLL5K Projektes und der Ergebnisse der Analysen aus SP1 und SP2 unternommen (AP3.2 und AP3.3). Ein wesentliches Ergebnis hier ist die Identifikation von *LRP1B* als potentieller Marker für einen Subtyp von seltenen B-Zell-Lymphomen, siehe Ergebnisse zu SP2.

Die Analysen auf Basis des MLL5K Datensatz fanden in enger Zusammenarbeit mit der Gruppe von Matthias Heinig statt, die Methoden aus dem maschinellen Lernen (ML) auf den Datensatz angewendet hat.

Beiträge zu diesem Teilprojekt durch das für uns bewilligte Personal beinhalteten:

- *Weitergehende Aufarbeitung des molekularen Rohdatensatzes:* Damit der MLL5K Datensatz als Input für die ML Modelle dienen konnte, wurde der aufgearbeitete Datensatz aus AP3.1 weitergehend prozessiert. Dies beinhaltete unter anderem die Filterung von technischen Artefakten und die Überführung in für die Modelle geeignete Formate.
- *Initiale Einteilung von Subtypen und Generierung von Überlebensdaten:* Eine bereits etablierte Subtypen-Klassifikation diente bei den ML Analysen entweder als Basis für die Modell-Erstellung („supervised“ Modelle) oder als Vergleichsdatsatz („unsupervised“ Modelle). Während der Laufzeit des Projekts veröffentlichte die WHO ihre aktualisierten Leitlinien zur Diagnose von hämatologischen Erkrankungen (WHO2022). Auf Basis der dort beschriebenen Subtypen wurden die MLL5K Patienten dementsprechend neu eingeteilt. Wir fokussierten uns hierbei auf die AML und MDS Kohorten, da diese die größte Probenanzahl hatten und somit auch die einzelnen Subtypen groß genug waren um eine statistisch sinnvolle Auswertung zu ermöglichen.
Des Weiteren wurden für alle analysierten Patienten aktuelle Überlebensdaten generiert, die als Teil der MLL-internen medizinischen Dokumentation zur Verfügung standen.
- *Biologische Interpretation der Ergebnisse:* Die Ergebnisse der ML Modelle wurden anschließend auf ihr Potential für eine Verbesserung der Klassifikation und Vorhersagekraft für das Überleben hin ausgewertet.

Im Folgenden werden die wesentlichen Ergebnisse der gemeinsamen Arbeiten zu den Arbeitspaketen im Rahmen von SP3 zusammengefasst:

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

In der initialen Phase des Projekts wurde der Datenbestand des 5.000 Genom-Projekts dahingehend aufarbeitet, dass er den Projektpartnern zur Verfügung gestellt werden konnte. Die Daten wurden in Form von gemappten BAM-Dateien, Varianten-Calls im VCF Format und Genexpressionsmatrizen bereitgestellt.

AP3.2: Subtypidentifizierung und –diagnose

Supervised Clustering der WHO2022-Subtypen in MDS und AML

Ziel der Analyse war es jeweils einen (multi-class) Klassifikator für die Erkennung der von der WHO2022 definierten Subtypen von AML (n=621) und MDS (n=698) 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)(p13.1;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. 8). Die Cluster sind durch AML-definierende genetische Veränderungen 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.

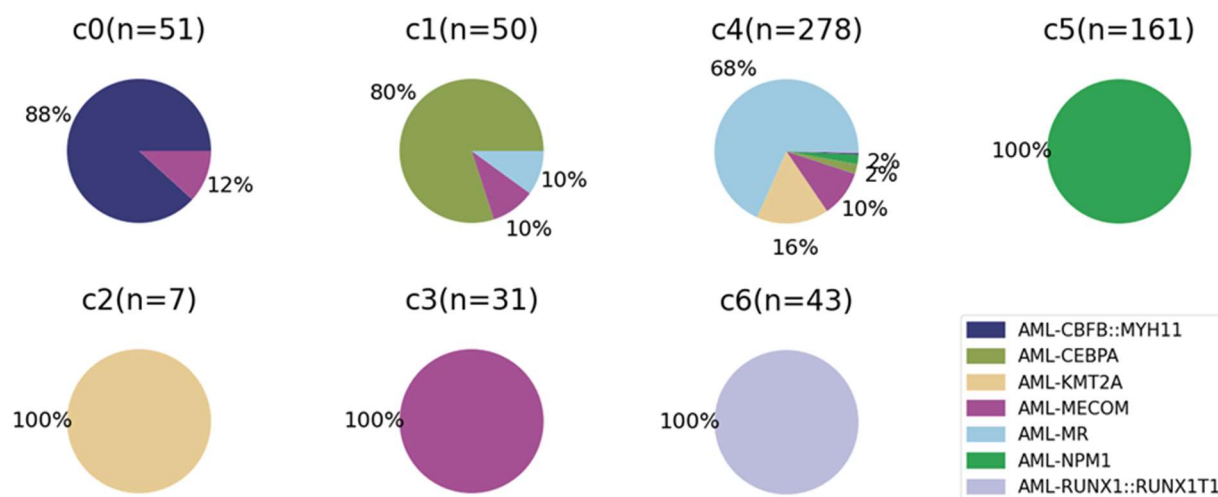


Abb. 8. Clustering der AML-Kohorte. Mittels GMM-Algorithmus wurden 7 verschiedene Cluster erzeugt. Farben spiegeln die jeweiligen WHO2022-Subtypen wieder.

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.

AP3.3: Integration 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-Regressionsmodell um die Zeit bis zum Todesfall in Abhängigkeit der Zugehörigkeit zu einem der Cluster zu modellieren.

In der AML Kohorte ergab die Überlebensanalyse einen C-Index von 0,6 (Abb. 9, links). Dieses Ergebnis ist dem Ergebnis der Überlebensanalyse auf Grundlage der WHO2022-Subtypen sehr ähnlich (C-Index von 0,62; Abb. 9, rechts). Wie Abb. 9 zu entnehmen ist, entsprechen die neuen Cluster größtenteils 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-rearrangierten AMLs. Es sind jedoch weitere Analysen notwendig, um diese Fälle genauer zu charakterisieren.

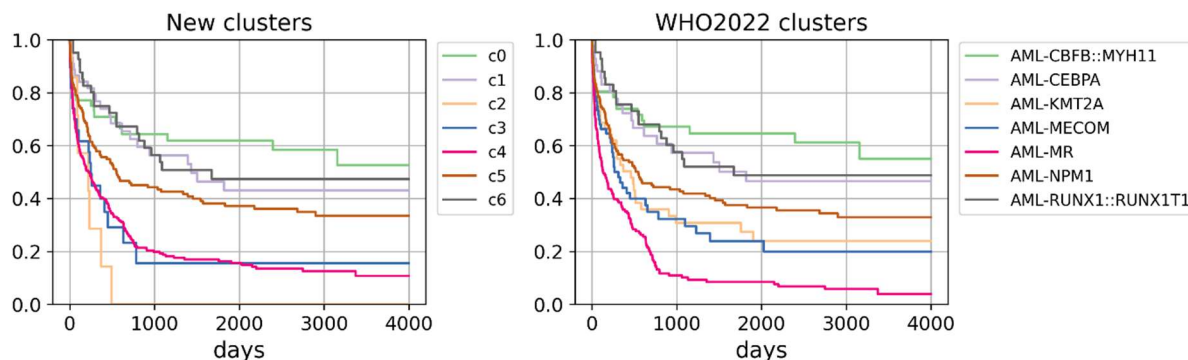


Abb. 9. Überlebensanalyse der AML-Kohorte. Überlebensanalyse der AML-Kohorte stratifiziert nach neuen Cluster (links) und nach WHO2022-Subtypen (rechts).

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 bei MDS 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. 2022). 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.

Literaturangaben

Bernard E, Tuechler H, Greenberg PL, Hasserjian RP, Arango Ossa JE, Nannya Y, Devlin SM, Creignou M, Pinel P, Monnier L, Gundem G, Medina-Martinez JS, Domenico D, Jädersten M, Germing U, Sanz G, van de Loosdrecht AA, Kosmider O, Follo MY, Thol F, Zamora L, Pinheiro RF, Pellagatti A, Elias HK, Haase D, Ganster C, Ades L, Tobinsson M,

Palomo L, Della Porta MG, Takaori-Kondo A, Ishikawa T, Chiba S, Kasahara S, Miyazaki Y, Viale A, Huberman K, Fenaux P, Belickova M, Savona MR, Klimek VM, Santos FPS, Boulwood J, Kotsianidis I, Santini V, Solé F, Platzbecker U, Heuser M, Valent P, Ohyashiki K, Finelli C, Voso MT, Shih LY, Fontenay M, Jansen JH, Cervera J, Gattermann N, Ebert BL, Bejar R, Malcovati L, Cazzola M, Ogawa S, Hellström-Lindberg E, Papaemmanuil E. Molecular International Prognostic Scoring System for Myelodysplastic Syndromes. *NEJM Evid.* 2022 Jul;1(7):EVIDoa2200008. doi: 10.1056/EVIDoa2200008. Epub 2022 Jun 12. PMID: 38319256.

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 May 20;16(1):70. doi: 10.1186/s13073-024-01331-6. PMID: 38769532; PMCID: PMC11103968.

Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, Gonzalez-Perez A, Lopez-Bigas N. A compendium of mutational cancer driver genes. *Nat Rev Cancer.* 2020 Oct;20(10):555-572. doi: 10.1038/s41568-020-0290-x. Epub 2020 Aug 10. PMID: 32778778.

Peramangalam PS, Surapally S, Veltri AJ, Zheng S, Burns R, Zhu N, Rao S, Muller-Tidow C, Bushweller JH, Pulikkan JA. N-MYC regulates cell survival via eIF4G1 in inv(16) acute myeloid leukemia. *Sci Adv.* 2024 Mar;10(9):eadh8493. doi: 10.1126/sciadv.adh8493. Epub 2024 Feb 28. PMID: 38416825; PMCID: PMC10901375.

Sakai S, Nakaseko C, Takeuchi M, Ohwada C, Shimizu N, Tsukamoto S, Kawaguchi T, Jiang M, Sato Y, Ebinuma H, Yokote K, Iwama A, Fukamachi I, Schneider WJ, Saito Y, Bujo H. Circulating soluble LR11/SorLA levels are highly increased and ameliorated by chemotherapy in acute leukemias. *Clin Chim Acta.* 2012 Oct 9;413(19-20):1542-8. doi: 10.1016/j.cca.2012.06.025. Epub 2012 Jun 29. PMID: 22750733.

Scheller IF, Lutz K, Mertes C, Yépez VA, Gagneur J. Improved detection of aberrant splicing with FRASER 2.0 and the intron Jaccard index. *Am J Hum Genet.* 2023 Dec 7;110(12):2056-2067. doi: 10.1016/j.ajhg.2023.10.014. Epub 2023 Nov 24. PMID: 38006880; PMCID: PMC10716352.

Wagner N, Çelik MH, Hölzlwimmer FR, Mertes C, Prokisch H, Yépez VA, Gagneur J. Aberrant splicing prediction across human tissues. *Nat Genet.* 2023 May;55(5):861-870. doi: 10.1038/s41588-023-01373-3. Epub 2023 May 4. PMID: 37142848.

Yépez VA, Mertes C, Müller MF, Klaproth-Andrade D, Wachutka L, Frésard L, Gusic M, Scheller IF, Goldberg PF, Prokisch H, Gagneur J. Detection of aberrant gene expression events in RNA sequencing data. *Nat Protoc.* 2021 Feb;16(2):1276-1296. doi: 10.1038/s41596-020-00462-5. Epub 2021 Jan 18. PMID: 33462443.

2. den wichtigsten Positionen des zahlenmäßigen Nachweises,

Position	Verwendung
Reise	Das Reisebudget wurde für das CompLS-Statusseminar (September 2023) im Marburg eingesetzt.
Personal	Es wurde eine Wissenschaftlerin zum 1.8.2021 (anstelle einst geplant 1.1.2021) eingestellt werden und der Beginn der Analysen erfolgte somit mit einigen Monaten Verspätung. Aufgrund dessen, und ähnlichen Verzögerungen bei den Projektpartnern was zu einer verspäteten Bereitstellung der zu analysierenden Daten führte, wird eine Verlängerung des Vorhabens um mehrere Monate angestrebt. Es wurde eine kostenneutrale Laufzeitverlängerung des Vorhabens bis 30.11.2023 bewilligt.

Die Mittel wurden wie geplant eingesetzt.

3. der Notwendigkeit und Angemessenheit der geleisteten Arbeit,

Die im Antrag für das Teilprojekt genannten Ziele und definierten Meilensteine wurden erreicht. Die beantragten Mittel für Reisen und Personal wurden benötigt und entsprechend eingesetzt.

4. dem voraussichtlichen Nutzen, insbesondere der Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans mit Zeithorizont,

Wirtschaftliche Erfolgsaussichten		Während der Laufzeit des Vorhabens	Im Anschluss an die Laufzeit
Ziel	Erläuterungen	<i>bitte zutreffende Felder ankreuzen</i>	<i>bitte zutreffende Felder ankreuzen</i>
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	<i>bitte zutreffende Felder ankreuzen</i>	<i>bitte zutreffende Felder ankreuzen</i>
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

Wissenschaftliche und wirtschaftliche Anschlussfähigkeit		Im Anschluss an das Vorhaben
Ziel	Erläuterungen	<i>Zeithorizont erläutern</i>
Wissenschaftliche, technische, strukturelle und versorgungsbezogene Verwertungsmöglichkeiten	Im Rahmen des Vorhabens wurden häufig Kandidatengene produziert, welche einer weiteren detaillierten Analyse bedürfen. Eine weiterführende Analyse in diesem Sinne wird am MLL in Kollaboration mit Piers Blombery (Universität Melbourne, Australien) durchgeführt, in der die SBLPN Kohorte genauer genetisch charakterisiert wird. Des Weiteren stellen die Ergebnisse Grundlagen für Anschlussprojekte dar, die das MLL zusammen mit den Arbeitsgruppen am Helmholtz München und der TU München eingegangen ist. Helmholtz: Hier werden Klassifikations-Modelle entwickelt, welche neben molekularen Daten auch weitere Datenmodalitäten (z.B. Blutbild) integrieren. TUM: Hier werden die Expressions- und Splice-Ausreißern zusammen mit genomischen Mutationen analysiert um Muster im gemeinsamen Auftreten dieser Mutationen zu identifizieren.	Seit Juli 2022 Seit August 2023 Seit Mai 2024

5. dem 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. den erfolgten oder geplanten Veröffentlichungen des Ergebnisses.

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 May 20;16(1):70. doi: 10.1186/s13073-024-01331-6. PMID: 38769532; PMCID: PMC11103968.