Genome Medicine

RESEARCH Open Access

Analysis of 3760 hematologic malignancies reveals rare transcriptomic aberrations of driver genes

Xueqi Cao[1,2], Sandra Huber[3], Ata Jadid Ahari[1], Franziska R. Traube[1,4], Marc Seifert[5], Christopher C. Oakes[6], Polina Secheyko[1,7], Sergey Vilov[8], Ines F. Scheller[1,8], Nils Wagner[1,9], Vicente A. Yépez[1], Piers Blombery[10,11,12], Torsten Haferlach[3], Matthias Heinig[1,8], Leonhard Wachutka[1*], Stephan Hutter[3*] and Julien Gagneur[1,2,8,13*]

Abstract

Background Rare oncogenic driver events, particularly affecting the expression or splicing of driver genes, are suspected to substantially contribute to the large heterogeneity of hematologic malignancies. However, their identification remains challenging.

Methods To address this issue, we generated the largest dataset to date of matched whole genome sequencing and total RNA sequencing of hematologic malignancies from 3760 patients spanning 24 disease entities. Taking advantage of our dataset size, we focused on discovering rare regulatory aberrations. Therefore, we called expression and splicing outliers using an extension of the workflow DROP (Detection of RNA Outliers Pipeline) and AbSplice, a variant effect predictor that identifies genetic variants causing aberrant splicing. We next trained a machine learning model integrating these results to prioritize new candidate disease-specific driver genes.

Results We found a median of seven expression outlier genes, two splicing outlier genes, and two rare splice-affecting variants per sample. Each category showed significant enrichment for already well-characterized driver genes, with odds ratios exceeding three among genes called in more than five samples. On held-out data, our integrative modeling significantly outperformed modeling based solely on genomic data and revealed promising novel candidate driver genes. Remarkably, we found a truncated form of the low density lipoprotein receptor *LRP1B* transcript to be aberrantly overexpressed in about half of hairy cell leukemia variant (HCL-V) samples and, to a lesser extent, in closely related B-cell neoplasms. This observation, which was confirmed in an independent cohort, suggests *LRP1B* as a novel marker for a HCL-V subclass and a yet unreported functional role of *LRP1B* within these rare entities.

Conclusions Altogether, our census of expression and splicing outliers for 24 hematologic malignancy entities and the companion computational workflow constitute unique resources to deepen our understanding of rare oncogenic events in hematologic cancers.

*Correspondence:
Leonhard Wachutka
wachutka@cit.tum.de
Stephan Hutter
stephan.hutter@mll.com
Julien Gagneur
gagneur@in.tum.de
Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Keywords Aberrant expression, Aberrant splicing, *LRP1B*, Hairy cell leukemia variant (HCL-V), Driver gene prediction

Background

Hematologic malignancies are characterized by abnormal blood cells in the bone marrow, peripheral blood, or lymphatic organs. They can occur in various forms, affecting the myeloid or lymphoid cell lineage. In 2020, hematologic malignancies accounted for approximately 2.5% of new cancer cases globally and accounted for 3.1% of cancer-associated mortality [1]. While some subtypes, like myeloproliferative neoplasm, exhibit a high degree of uniformity in their manifestation and genetic profile, others, like myelodysplastic neoplasm, display a significantly broader spectrum, hampering correct diagnosis and therapy decisions, which negatively impacts treatment outcomes and survival [2]. Thus, better understanding the variety of oncogenic events for each disease entity is of utmost interest to refine diagnostics and facilitate the development of new therapeutic options.

Within the last decade, the identification of driver genes in hematologic malignancies has been dramatically enhanced. To this end, functional screens such as CRISPR [3, 4] and transposon screens [5] on model systems have been applied. Complementary to these efforts, next-generation sequencing analyses of primary clinical samples [6, 7] were employed, which better capture the in vivo biology. This research has provided valuable insights into the underlying genetic landscape of each entity and triggered a revision of the classification systems, which now emphasize genomics-based categorization of various leukemia and lymphoma entities [2, 8–10]. However, despite significant progress in understanding recurrent driver mutations in hematologic malignancies, much remains to be learned about the rare events within each disease entity that drive their individual development and progression [11, 12]. Such rare events could arise not only from somatic mutations but also from rare germline variants, as supported by an increasing number of studies unraveling the implication of rare genetic predispositions to cancer [13–18].

Alterations of gene expression and splicing play a key role in cancer [19], particularly in hematologic malignancy pathogenesis [20–30]. For instance, many hematologic malignancies are characterized by altered expression resulting from gene rearrangements that lead to the overexpression of specific transcription factors or cell cycle regulators. Examples include acute myeloid leukemia (AML) with defining genetic alterations [31], *BCR::ABL1*-positive chronic myeloid leukemia (CML) [32], or *CCND1* rearrangements in mantle cell lymphoma (MCL) [2, 33]. Moreover, aberrant splicing can generate gain or loss-of-function transcript isoforms of driver genes for many cancer types [34]. For example, multiple aberrant splice isoforms of the *TP53* transcript have been observed in CML, even in the absence of genomic mutations around exon–intron junctions [35]. However, a systematic analysis of rare expression and splicing aberrations among hematologic malignancies is still lacking.

To address this gap, we conducted a comprehensive analysis of genomes (whole genome sequencing, WGS) and matched transcriptomes (total RNA sequencing, RNA-Seq) of tumor tissues from 3760 patients spanning 24 hematologic malignancy entities (Fig. 1). We analyzed this data using RNA-seq-based expression and splicing outlier callers, as well as AbSplice [36], a tool we recently published that predicts rare genetic variants causing aberrant splicing. We demonstrate how these results can be utilized to identify a novel marker for a rare entity and enhance the prediction of hematologic malignancy driver genes beyond the commonly used mutational recurrence. In summary, our study aims to deepen our understanding of the role of rare gene expression and RNA splicing in the development of hematologic malignancies and provide novel driver gene candidates.

Methods

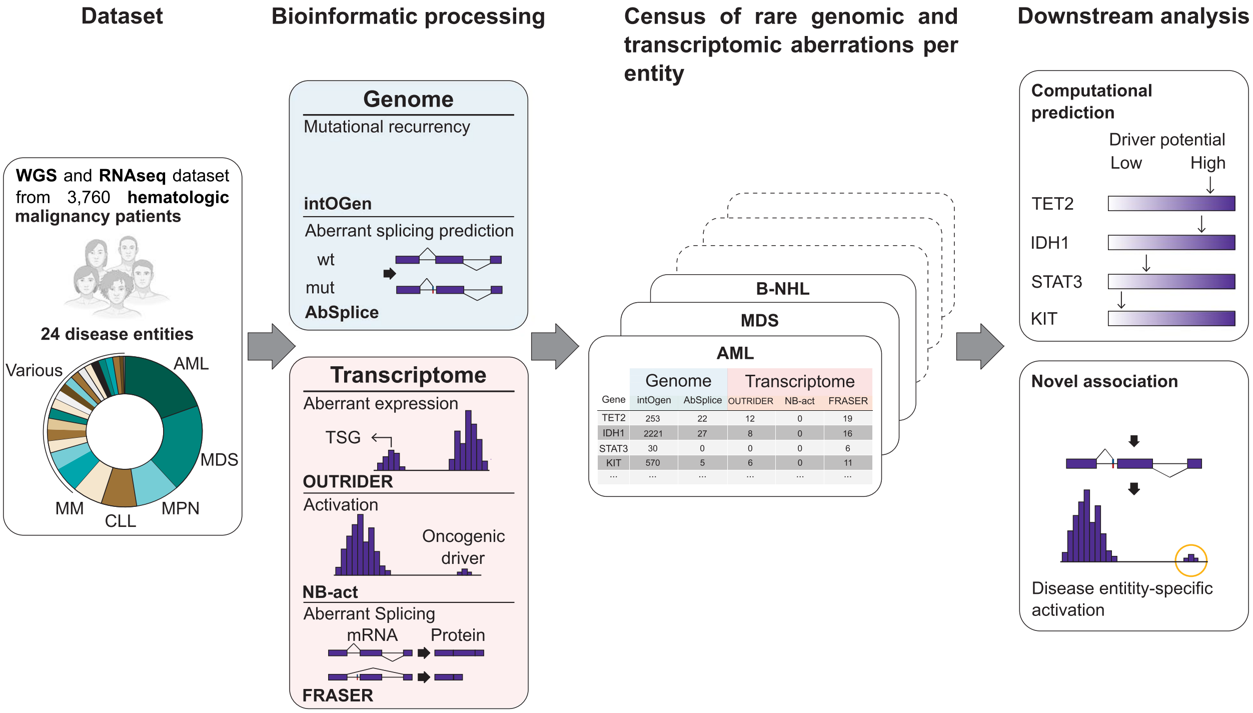
Dataset

Patients

We used genomic and transcriptomic data from the Munich Leukemia Laboratory (MLL). We included a total of 3760 tumor samples sent to the MLL between September 2005 and April 2019 for routine diagnostic workup (Table 1). Diagnoses from peripheral blood or bone marrow were based on cytomorphology, immunophenotype, cytogenetics, and molecular genetics, as previously described [40–42]. All patients or their legal guardians gave written informed consent for genetic analyses and to the use of laboratory results and clinical data for research purposes, according to the Declaration of Helsinki. The study was approved by the MLL's institutional review board. The dataset spanned 24 disease entities (Table 1, Table S1).

Sample preparation

DNA and total RNA from peripheral blood and bone marrow samples were extracted using the MagNA Pure 96 Instrument and the MagNAPure96 DNA and Viral NA LV Kit and MagNA Pure 96 Cellular RNA LV Kit, respectively (Roche LifeScience, Mannheim, Germany). WGS and RNA-seq were performed on the prepared samples (Supplementary Materials and Methods).



	Genome		Transcriptome		
Gene	intOgen	AbSplice	OUTRIDER	NB-act	FRASER
TET2	253	22	12	0	19
IDH1	2221	27	8	0	16
STAT3	30	0	0	0	6
KIT	570	5	6	0	11
...	...	...	...	...	...

Fig. 1 Overview of the study. Dataset: Whole genome sequencing and total RNA sequencing of 3760 hematologic malignancies spanning 24 different disease entities. Bioinformatic processing: On genomic data, IntOGen captures recurrent mutational patterns [37], and AbSplice predicts variants causing aberrant splicing [36]. Working on RNA-seq data, OUTRIDER calls expression outliers of commonly expressed genes [38], NB-act calls overexpression of rarely expressed genes (Methods), and FRASER calls splicing outliers [39]. Census: a unique collection of genomic and transcriptomic aberrations for 24 hematologic malignancy entities. Downstream analysis: driver gene prediction and enrichment analysis per disease entity

Variant calling and annotation

Variant calling of single-nucleotide variants, short insertions and deletions, structural variants, copy number variations, and gene fusions were performed on all samples as described previously [43–46] (Supplementary Materials and Methods). The analysis was based on the GENCODE v33 [47] annotation and using the reference genome GRCh37.

One particularity of our setting is that matched healthy control tissues are rarely available since it is generally not required for diagnosis during routine diagnostics. Therefore, we filtered variants based on sex-matched normal samples (Supplementary Materials and Methods). Hence, these variants encompass both somatic and germline variants. While rare germline variants can include variants predisposing to leukemia, it is less likely for common germline variants. To discard common germline variants and reduce artifacts from the data, we filtered single-nucleotide variants, short insertions, and deletions with the following criteria ("unmatched_filter2", see Supplementary Results):

(1) Only consider variants emitted with "PASS" quality
(2) Discard variants with gnomAD v2.1.1 [48] minor allele frequency > 0.05% to discard common germline variants
(3) Discard variants with sample variant allele frequency < 0.1, as we found an excess for those low VAF variants, perhaps as a consequence of DNA acoustic shearing [49]

Analogously, we filtered structural variant calls with the following criteria:

(1) Only consider variants emitted with "PASS" quality
(2) Only consider variants if they have three or more paired-reads supporting it
(3) Discard variants found with exact breakpoint locations in the gnomAD SV database [50]
(4) Discard variants that are found in four or more different myeloid and four or more different lymphatic entities

Altogether, our variant filtering procedures yielded a combination of somatic and rare germline variants.

Table 1 Disease entities enrolled and corresponding study groups

Disease entity	Abbreviation	Number of samples per disease entity	Study group	Number of samples per study group
Acute myeloid leukemia	AML	730	AML	730
Myelodysplastic neoplasm	MDS	713	MDS	713
Marginal zone lymphoma	MZL	71	MatureB group	375
Mantle cell lymphoma	MCL	71		
High-grade B-cell lymphoma	HGBL	61		
Lymphoplasmacytic lymphoma	LPL	59		
Follicular lymphoma	FL	57		
Other B-cell non-Hodgkin lymphomas	B-NHL	56		
Myeloproliferative neoplasm	MPN	344	MPN	344
Chronic lymphocytic leukemia	CLL	279	CLL	279
Multiple myeloma	MM	236	PCN group	247
Monoclonal gammopathy of undetermined significance	MGUS	11		
Myelodysplastic/myeloproliferative neoplasm, unclassifiable	MDS/MPN-U	81	MDS/MPN group	207
Atypical chronic myeloid leukemia	aCML	67		
Myelodysplastic/myeloproliferative neoplasm with ring sideroblasts and thrombocytosis	MDS/MPN-RS-T	59		
B-cell precursor acute lymphoblastic leukemia	BCP-ALL	204	BCP-ALL	204
T-cell non-Hodgkin lymphoma	T-NHL	90	T-cell group	174
T-cell acute lymphoblastic leukemia	T-ALL	84		
Chronic myelomonocytic leukemia	CMML	142	CMML	142
Mastocytosis		98	Mastocytosis	98
Hairy cell leukemia	HCL	68	Hairy-cell group	97
Hairy cell leukemia variant	HCL-V	29		
Chronic myeloid leukemia	CML	92	CML	92
Chronic lymphoproliferative disorder of natural killer cells	CLPD-NK	58	NK	58

To predict the effects of variants on splicing, we applied AbSplice-DNA v1.0.0 [36] to the filtered variants. AbSplice-DNA estimates the probability that a rare variant causes aberrant splicing in a given tissue, which takes the variant and splice annotations, so-called SpliceMaps, as input. We generated SpliceMaps for each study group as described previously [36]. The variants with an AbSplice-DNA prediction score ≥ 0.2 were classified as splice-affecting variants. Further variant effect predictions were obtained by applying Ensembl VEP v81 [51]. VEP splice-related variants were defined as VEP calculated consequences "splice_acceptor_variant" (2 base region at the 3′ end of an intron), "splice_donor_variant" (2 base region at the 5′ end of an intron), and "splice_region_variant" (1–3 bases of the exon or 3–8 bases of the intron).

Gene-level mutational recurrence scores were obtained using IntOGen commit 437a047 [37], and MutSigCV v1.41 [52]. The IntOGen framework has a hard-coded cutoff of 10,000 on the total number of variants per sample. This cutoff has been set from studies based on somatic variants only. In our setting, which includes an excess of rare germline variants, the cutoff is too low. However, within the IntOGen framework, the input of all 7 individual tools is restricted to variants from coding transcripts only. Therefore, to circumvent IntOGen's rejection of samples with a large number of variants, we have filtered for variants on UTRs and coding sequences prior to running IntOGen using the VEP calculated impact 'HIGH', 'MODERATE', 'LOW', or calculated consequence "5_prime_UTR_variant," "3_prime_UTR_variant," and "coding_sequence_variant."

Moreover, the variants were further filtered using the following criteria, as these more stringent filters led to improved driver gene enrichments of IntOGen ("unmatched_filter3", see Supplementary Results):

(1) Discard variants with sample variant allele frequency < 0.15
(2) Only consider variants supported by at least 20 reads

Expression and splicing outlier calling

We implemented a Python backend version of OUTRIDER v.1.99.0 [38] for scalability (see Availability of data and materials) and applied it to RNA-seq data using DROP v1.1.0 [53] default settings to call expression outliers at a false discovery rate (FDR) of 0.05. Expression outliers with lower-than-average expressions (z-score < 0) were defined as underexpression outliers, and higher-than-average expressions (z-score > 0) as overexpression outliers.

FRASER v.1.99.0 [39] was applied to RNA-seq data using DROP v1.2.3 [53] with default settings to call splicing outliers (FDR < 0.05). We grouped the samples into 14 study groups based on hematopoietic cell origins and pathologies [2, 8–10] to ensure sufficient sample sizes (Table 1, N at least 58). Splicing outliers with delta Intron Jaccard Index > 0 were defined as overrepresented splicing outliers, and delta Intron Jaccard Index < 0 as underrepresented splicing outliers.

As OUTRIDER is restricted to commonly expressed genes defined as genes with fragments per kilobase of transcript per million mapped reads (FPKM) larger than 1 in at least 5% of the samples, we introduced a novel method to detect aberrant activation of genes usually not expressed. This method, NB-act (Negative Binomial activation), provides p-values for observed fragment count (read pairs) for each gene in each sample under the null hypothesis that the gene is not expressed in the sample. Specifically, NB-act computes the probability of observing a certain number of fragments or more for a gene in a particular sample, assuming a negative binomial distribution with an expected baseline expression of 1 FPKM and a dispersion parameter of 0.02 (Supplementary Materials and Methods). The FPKM value of 1 corresponds to the threshold separating expressed from non-expressed genes in OUTRIDER [38]. The dispersion parameter of 0.02 corresponds to the empirically observed lowest dispersion values estimated by OUTRIDER on expressed genes. As low dispersion corresponds to high variance, we chose a low dispersion value for NB-act to be conservative. NB-act was applied to rarely expressed genes (FPKM > 1 in at least one sample but less than 5% of the samples).

Enrichment for driver gene and variant categories

The enrichment of cancer driver genes was evaluated by the Fisher test using Cancer Gene Census (CGC) GRCh37 v97 [54] (Supplementary Materials and Methods). The role of hematologic malignancy driver genes was determined using annotation "Tissue Type" and "Role in Cancer".

As promoter variants, we considered all single-nucleotide variants and short insertions and deletions less than 2000 bp away from the transcription start site of the corresponding gene. Frameshift and stop-gain variants were detected with Ensembl VEP v81 [51]. For copy number variation, we considered a gene affected by copy number variation if its position overlaps with the "+" or "−" regions called by GATK CallCopyRatioSegments [55]. The enrichment analysis of variants within expression outlier gene-sample pairs was performed using the Fisher test. Sample-gene pairs with other variants were excluded apart from those for which enrichment was calculated.

Variance component analysis

The variance component analysis was conducted using a linear model. We used the logarithmized copy ratio and a binarization of the rare VEP high-impact, rare promoter, and rare structural variants affecting a gene-sample combination as the independent variables. As for the dependent variable, we used an autoencoder corrected expression zScore for each gene-sample combination. The linear model was fitted for every gene individually, followed by ANOVA. The variance explained by each independent variable was calculated as the sum of square errors of each independent variable over all independent variables. The resulting variance components were normalized to add up to one, and we reported the mean value.

Survival analysis

Overall survival (OS) analyses were performed according to Kaplan–Meier and compared using two-sided log-rank tests with the statistical software R v4.2.2 with the survival [56] and survminer [57] packages. The overall survival was calculated as the time from diagnosis to death or last follow-up.

Driver gene prediction

Machine-learning models were trained to predict the probability of a gene to be a hematologic malignancy driver gene. To this end, random forest classifiers (Python package scikit-learn v1.0.2) [58] were trained to predict driver genes using gene-level features obtained from all samples on the one hand and each of the 14 study groups on the other hand. We also fitted logistic regression, XGboost, and fully connected neural networks. The gene-level features consisted of 21 features from seven gene-level metrics of seven IntOGen tools, nine features from AbSplice-DNA scores, 22 features from OUTRIDER obtained using combinations of fold-change direction, significance, and effect size cutoffs, and, similarly, 11 features from NB-act and 22 features from FRASER (Supplementary Materials and Methods). For those models integrating external gene functional data, co-essential modules from DepMap [59] and 256-dimensional functional gene embeddings [60] were further included as features. In total, 377 genes listed among the hematologic panel genes (Supplementary Materials and Methods) or the hematologic malignancy driver genes from CGC GRCh37 v97 [54] were used as the positive class for the classifiers. The random forests were trained using the function "RandomForestClassifier" with the minimum number of samples required to split an internal node set to 19, the maximal tree depth set to 10, and default settings otherwise (Supplementary Materials and Methods). The models were trained with fivefold cross-validation, stratified to preserve the percentage of positive class in all folds, using the function "StratifiedKFold" from the Python package scikit-learn v1.0.2 [58]. Performances were evaluated by precision-recall curves on the validation data. This process was repeated 10 times (random repeats) to assess the variability of the performance estimates. Subsequently, only prediction values based on validation data were used for candidate curation. As the predicted probabilities range very differently across study groups, no fixed threshold was set, and we focused on the rank of the predicted genes. The list of approved drugs was obtained from the Pharos database (https://pharos.nih.gov/) [61]. The benchmark against IntOGen and MutSigCV was conducted on all study groups.

Further detailed descriptions are available in the Supplementary Materials and Methods.

Validation HCL-V dataset

Patients

The validation HCL-V dataset included a total of 42 patients, including 14 HCL-V and 28 HCL patients. The HCL-V patients were diagnosed based on immunophenotype and morphological characteristics consistent with HCL-V, including all CD5 negative, CD11c positive, and CD123 negative markers. All HCL-V patients were confirmed as *BRAF*-V600E negative.

Sample preparation

Tumor cells were isolated by cell sorting (purity > 99%) as FSC^{high}CD20^{+}CD11c^{+} and kappa/lambda light-chain restriction. RNA was isolated by the RNeasy Micro Kit (Qiagen, Hilden, Germany).

RNA sequencing, read mapping

For the generation of RNA-sequencing libraries, the NuGEN Trio RNA-Seq System (NuGEN, Redwood City, California) was used. Samples were split equally and processed in independent sequencing steps to allow for the correction of batch effects. Sequencing was performed with paired-end sequencing and two times 100-bp length. Sequences were aligned with HiSAT2 v2.1.0 [62] to the GRCh38.

Gene expression analysis

Gene expression was analyzed with DESeq2 [63]. The size factor was calculated using DEseq2. The counts were normalized by size factors. The activation of *LRP1B* was determined with a two-component Gaussian mixture clustering with equal variance on size factor-normalized and log-transformed counts, using the R package "mclust" v.6.0.0 [64].

Results

We investigated WGS and RNA-seq data from 3760 tumor samples representing 24 different types of leukemia and lymphoma (Table 1). This dataset has been collected from routine diagnostics and is the largest collection of hematologic malignancy samples with WGS and matched RNA-seq, which also includes rare disease entities like hairy cell leukemia variant (HCL-V) and chronic lymphoproliferative disorder of natural killer cells. In order to restrict our analysis to putative rare germline and somatic variants, we filtered variants called on WGS with stringent quality filters and population allele frequency to discard artifacts and common germline variants, respectively (Table S2, S3 and S4) [48]. We next annotated the genes using the seven features from the software IntOGen, which include positional recurrence of variants in genome sequence (OncodriveCLUSTL), positional recurrence of variants in protein conformation (HotMAPS), enrichment of variants in functional domains (smRegions), three alternative measures of selection strength inferred from synonymous and non-synonymous variants (CBaSE, MutPanning, and dNdScv), and OncodriveFML, a method identifying excess of variants across tumors in both coding and non-coding genomic regions [37, 65–71]. Moreover, we annotated genetic variants falling into gene bodies, including deep intronic variants, with AbSplice-DNA, a tool predicting variants causing aberrant splicing [36]. On the RNA-seq data, we used OUTRIDER on a total of 12,966 protein-coding genes commonly expressed across the dataset to call high or low expression outliers, and FRASER to call splicing outliers [38, 39]. We also introduced a new method, NB-act, to call rare aberrant activation of genes mainly not expressed (Methods). As summarized in Table 2, these methods provide qualitatively complementary evidence for detecting and predicting driver genes. Combining all these results, we established a unique census of genomic and transcriptomic aberrations in 3760 hematologic malignancy samples (Fig. 1).

Expression outliers are enriched for hematologic malignancy driver genes

Case–control differential expression analysis identifies common differences between tumor and healthy samples

Table 2 Complementary of the methods used as input for the integrative analyses

Input	Method	Goal	Limitation
Genetic variants	IntOGen	• Detect different types of the recurrence of genomic alterations in genes • Combine seven tools that cover multiple aspects of cancer driver gene detection	• Focus only on genetic variants • Focus only on single nucleotide variants and short indels, while structural variants, epigenetic silencing events, and germline susceptibility variants are not considered
Genetic variants	AbSplice	• Estimates the probability for a genetic variant to cause aberrant splicing • Integrates deep learning sequence-based models (SpliceAI and MMSplice) with quantitative maps of splicing levels in tissues of interest (SpliceMap) • It can be used to trace RNA-seq-based aberrant splicing calls back to the genomic-level variant	• For deep intronic variants, AbSplice performs not as well as near splice site variants • SpliceMaps need to be created if new tissue or cell types are added
RNA-seq	OUTRIDER	• Detects RNA expression outlier, independently of genetic variants • Accounts for covariations using denoising autoencoder	• Applies only to genes typically expressed in the considered cohort. Fails at calling activation of otherwise not expressed genes • Sufficiently large cohort is required (> 60 samples) to detect outliers reliably
RNA-seq	NB-act	• Detects aberrantly activated genes in RNA-seq data, which complements OUTRIDER	• In comparison to underexpression outliers for which NMD-triggering variants provide orthogonal ground truth, benchmarking data based on rare variant annotation is less certain for gene activation
RNA-seq	FRASER	• Detects aberrantly spliced genes in RNA-seq data • Accounts for sources of covariation using denoising autoencoder • Intron-centric: no prior annotation needed, and does not require building clusters of introns sharing splice sites, which can get prohibitively big and lead to modeling complications	• Sufficiently large cohort is required (> 50 samples) to detect outliers reliably • Can overlook some genuinely pathogenic isoforms, especially rapidly degraded splice isoforms

and, therefore, can be suited to identify recurrently differentially expressed genes in tumors. To identify potential rare driver genes, we instead employed expression outlier analysis, which calls rare, aberrantly high, or low expression of a gene within a dataset. OUTRIDER calls expression outliers by modeling read count distribution across samples and reporting the statistical significance of extreme observations. Moreover, OUTRIDER controls for gene expression covariations, which allows automatic corrections for technical sources of variation, such as batch effects, and adjusts for transcriptome-wide co-regulation patterns due to trans-acting regulatory changes. Deviations from these variations have been shown to be enriched for rare variants with strong cis-regulatory effects [38] and help identify causes of rare disorders [72–75]. Applying OUTRIDER, we called 21,264 underexpression outliers (median of 2 per sample) and 14,041 overexpression outliers (median of 2 per sample) on 10,193 different protein-coding genes (Figure S1-S2, Table S5 and S6). To verify whether these outliers are associated with cancer, we calculated their enrichment for reported hematologic malignancy driver genes, which we adapted from CGC [54]. We observed a strong enrichment for CGC hematologic tumor suppressor genes among underexpression outliers and for CGC hematologic oncogenes among overexpression outliers (Fig. 2A, B). Notably, the genes called as outliers in more than five samples exhibited the highest enrichment, indicating that genes frequently called as outliers are more likely to be oncogenic. Moreover, we found that the number of expression outliers per sample was unevenly distributed (Figure S1-S2). Samples with numerous outliers may either represent cases where OUTRIDER could not adequately fit the data or situations where the gene regulatory network is globally affected, resulting in widespread expression aberrations throughout the genome. We reasoned that the enrichment for driver genes among expression outliers could be lower in all those samples. Indeed, the enrichment for tumor suppressor genes and oncogenes increased when focusing on the three most significant outliers (at most three) in each sample (Fig. 2A, B).

OUTRIDER filters out genes expressed in less than 5% of samples due to statistical modeling limitations, leaving a gap in detecting rare gene activation. To fill this gap, we developed a complementary algorithm, NB-act (Methods). We applied it to the 6017 rarely expressed protein-coding genes filtered out by OUTRIDER. NB-act identified 10,263 activation outliers among 1623 genes (with a median of 0 and 75% quantile of 2 per sample, Figure S3, Table S7). We observed a notable enrichment for CGC hematologic oncogenes among all activation

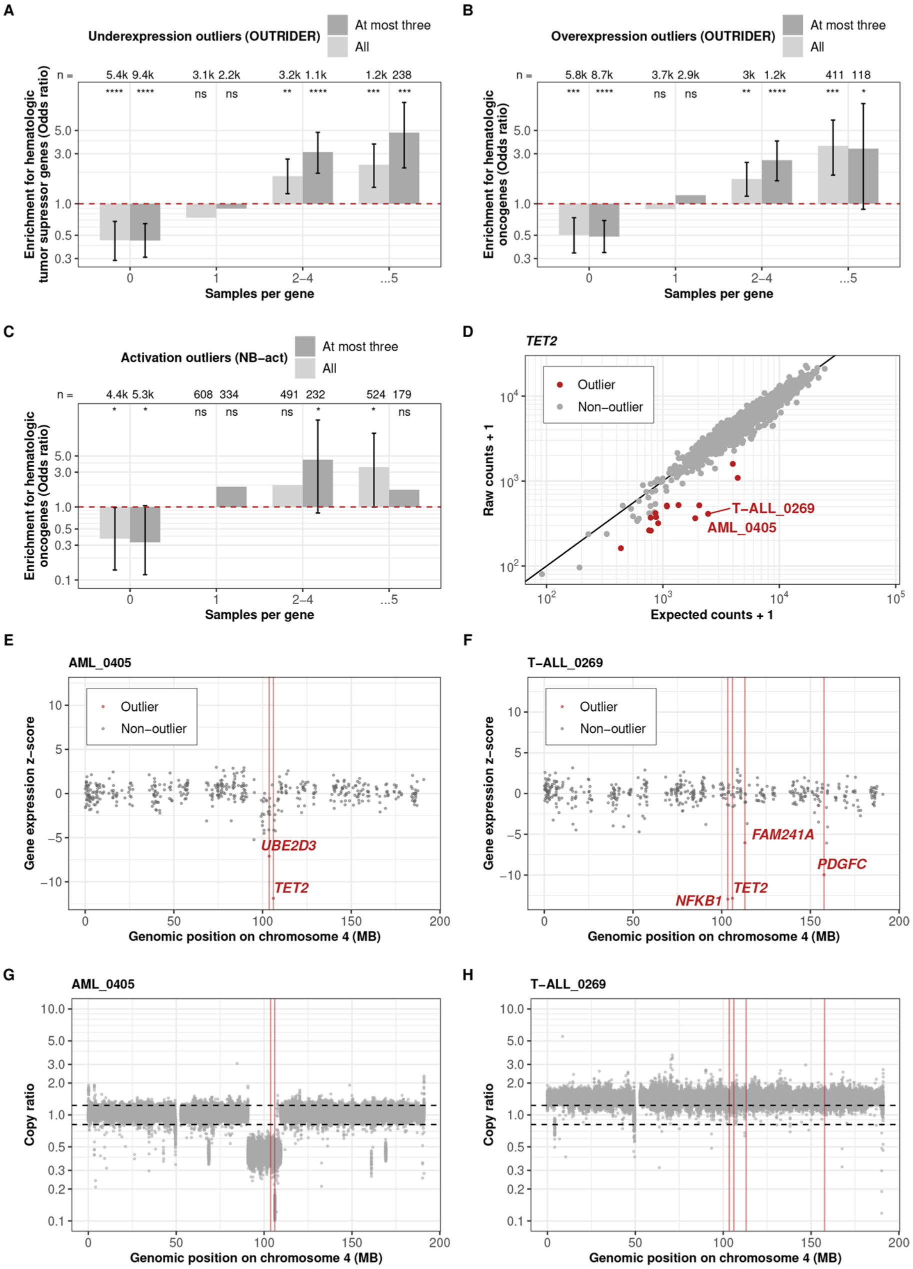


Fig. 2 Expression outliers are enriched for hematologic malignancy driver genes. **A** Enrichment for CGC hematologic tumor suppressor genes among all genes called by OUTRIDER as well as at most three significant genes per sample that were called to be underexpression outliers. The genes are stratified by the number of samples in which the gene is called as an outlier. Numbers of the genes and nominal significances from the Fisher test are labeled at the top of the bars (ns: not significant; *: $P \leq 0.05$; **: $P \leq 0.01$; ***: $P \leq 0.001$; ****: $P \leq 0.0001$). Error bars of the odds ratio (95% confidence intervals) are shown where the Fisher test is significant. **B**, **C** As in **A** for CGC hematologic oncogenes among overexpression and activation outliers, respectively. **D** Raw RNA-seq read counts per sample against expected counts for *TET2*. For fifteen samples (red), *TET2* was called as an underexpression outlier, including samples AML_0405 and T-ALL_0269, showcased in the panels (**E–H**). **E**, **F** Gene expression *z*-score among all samples against the genomic position of the genes on chromosome 4. **G**, **H** Copy ratio within chromosome 4 binned into 1-kb region. Red vertical lines mark the genomic position of the expression outliers. Black dashed lines mark the estimated region of no copy number variation. The underexpression outliers in *TET2* and *UBE2D3* in sample AML_0405 reflect the consequence of copy number loss that is very specific to the *TET2* locus. In contrast, an extra copy of whole chromosome 4 (karyotype: 47,XY,+4) is found in sample T-ALL_0269. *TET2* reduced expression must have a different cause in this sample

outliers (Fig. 2C). Here too, restricting to at most three outliers per sample increased the enrichment. Altogether, these analyses provide a unique set of aberrantly expressed genes in hematologic malignancies with strong enrichment for driver genes.

We next investigated how expression outliers were associated with genomic alterations. We observed a strong enrichment of relevant genomic aberrations among the expression outliers. Overall, 22.2% of the underexpression outliers overlapped a copy number loss, and 14.0% of the overexpression outliers and 12.6% of the activation outliers overlapped a copy number gain. We also performed a variance component analysis of OUTRIDER-corrected expression levels and found that copy ratio explained the largest part of the variance explained by genomic alterations (typically 60%, Figure S4-S5), consistent with a predominant role of copy number alterations in transcriptomic aberrations in cancer [76, 77]. We also found significant enrichments for rare variants associated with nonsense-mediated decay (stop-gained, frameshift, and splice-related) as well as structural variants and variants found in the promoter region among underexpression outliers (Methods, Figure S6), consistent with earlier reports in non-cancer samples [78], and with genomic alterations underpinning allele-specific expression in cancer [19]. Structural variants were also significantly enriched among overexpression and activation outliers (Figure S7-S8). Interestingly, we found enrichments for rare splice-related variants and rare frameshift variants among overexpression outliers but not activation outliers, perhaps because, in some cases, these variants lead to RNAs with increased stability (Figure S7-S8). Overall, these enrichments for relevant genomic aberrations support the reliability of the expression outlier calls and provide a genetic explanation for a substantial fraction of them.

Investigations of expression outlier events of the tumor suppressor gene *TET2* illustrate how this catalog can be used. Loss of function of *TET2* has been reported in myeloid leukemia due to splice site mutations, out-of-frame insertions or deletions, and base substitutions [79–82]. We found 15 underexpression outlier events (two in AML, one in myelodysplastic neoplasm, seven in B-cell precursor acute lymphoblastic leukemia, and five in T-cell acute lymphoblastic leukemia) for *TET2* across all samples and no overexpression outliers (Fig. 2D), consistent with its role as a tumor suppressor gene. To gain further insights, we showcased the two samples with the lowest fold changes (samples AML_0405 and T-ALL_0269). For these two samples, *TET2* was among the most extreme outliers (Figure S9-S10). However, neither the single nucleotide variants, short insertions and deletions, structural variants, nor gene fusions explained the observed decrease in *TET2* expression. Inspection of the genomic coverage indicated a loss of the *TET2* locus nested within a single-copy loss of a larger region in chromosome 4 for sample AML_0405 likely explaining the underexpression (Fig. 2E, G). This explanation was further supported by the decreased expression of the neighbor gene *UBE2D3* (Fig. 2E). In contrast, genomic coverage investigations did not provide an explanation for the reduced expression of *TET2* in sample T-ALL_0269 (Fig. 2F, H). Further outliers, including *TET2* in sample T-ALL_0269, could reflect yet-to-be-interpreted genetic variants, epigenetic causes, or outlier caller false positives [83].

In summary, both OUTRIDER and NB-act reveal aberrantly expressed genes enriched for driver genes and can be used to identify downregulated tumor suppressor genes or activated oncogenes in individual samples.

Rare splicing aberrations are enriched for hematologic malignancy driver genes

We applied FRASER to detect aberrant splicing events (aberrant usage of existing or novel splice sites) within our samples, which could be caused by events such as alternative exon usage, intron retention, alternative donor or acceptor site usage, usage of deep intronic donor and acceptor sites, or truncation of parts of the transcript. Like OUTRIDER for expression, FRASER is a tool to call splicing outliers while controlling for covariations, a task that is distinct from calling differential splicing between groups [39]. We called 43,464 splicing outliers across 35,410 gene-level splicing outlier events on a total of 7591 genes in 2854 samples (Figure S11, Table S8). Remarkably, we observed a substantial enrichment for CGC hematologic tumor suppressor genes among these splicing outliers (Fig. 3A). As for expression outliers, restricting to at most three outliers per sample increased the enrichment. Moreover, the genes called as splicing outliers in more than five samples exhibited the highest enrichment, suggesting a higher potential of being oncogenic.

Calling splicing aberrations from RNA-seq data can point towards mutations that have been overlooked using the genomic data alone. As an example, we showcase *RB1*, a well-known tumor suppressor gene, which can be inactivated by various mechanisms, including intragenic mutations, methylation of the promoter region, and chromosomal deletions [84]. In sample B-NHL_1747, we identified an unusual exon-skipping event of the 20th exon in RB1 (Figure S12). Analyzing the WGS data of this sample revealed a deletion of 118-base-pair in the region of the 20th exon, which led to a frame-shift exon-skipping (Figure S13). However, this structural variant was initially discarded in our variant calling pipeline due to stringent filtering, where we required paired-read support for considering a variant, while this deletion was only supported by split-reads. While anecdotal, this example is reminiscent of a similar case observed in rare diseases, where expression outlier helped to identify a promoter variant responsible for underexpression, while the variant was initially not prioritized during whole exome sequencing analysis due to its location [85]. Using FRASER, which accurately captured the consequence of this initially discarded structural variant as a splicing outlier, we were now able to recognize the significant transcriptomic impact of this variant and rescued it.

As a complementary approach to RNA-seq-based splicing outlier calling, we considered AbSplice [36], a recently published algorithm that predicts whether a rare variant causes aberrant splicing. Here, we applied it for the first time to hematologic malignancy samples. AbSplice integrates two sequence-based machine learning tools, MMSplice [86] and SpliceAI [87], with so-called SpliceMaps, which are quantified tissue-specific usages of splice sites, including non-annotated and weak splice sites. We derived SpliceMaps from the raw RNA-seq data (Methods). AbSplice classified 7160 rare variants as splice-affecting out of 275,899,978 pre-filtered variants, resulting in 7074 genes predicted to be affected across 3093 samples (with a median of 2 per sample, Figure S14, Table S9 and S10). These results demonstrated a strong enrichment for CGC hematologic tumor suppressor genes among all genes carrying splice-affecting variants (Fig. 3B). Furthermore, when focusing on the genes that were predicted to have recurrent aberrant splicing events across multiple samples, we observed an even higher enrichment, indicating a tendency for recurrent splice-affecting variants to occur in known hematologic malignancy driver genes (Fig. 3B). In contrast to the RNA-seq-based splicing outlier calls, restricting AbSplice predictions to at most three calls per sample did not lead to a significantly higher enrichment, likely due to the limited number of samples with a high number of variants displaying strong AbSplice scores. Collectively, these results indicate that the rare splicing aberrations predicted from the genome can contribute to identifying potential driver genes in hematologic malignancies.

Overall, 15% of the AbSplice predictions (1049 out of 7089, Fig. 3C) were also called as splicing outliers by FRASER, which is lower but not far off the claimed precision of AbSplice at the cutoff we used (predicted precision cutoff = 0.2). This proportion of FRASER splicing outliers was much larger using AbSplice than when using VEP splice-related rare variants (1.2%, 1504 out of 121,256, Fig. 3C). Among those 1049 gene-sample pairs common to AbSplice and FRASER, the enrichment for CGC hematologic tumor suppressor genes was substantially stronger than when considering AbSplice predictions or FRASER calls separately (Fig. 3D). We also found significantly high enrichments for VEP splice-related rare variants and AbSplice variants among splicing outliers (Methods, Figure S15 and S16). As an example of an event both predicted by AbSplice and called by FRASER, we showcase *CD79A*, a well-known oncogene frequently affected by somatic mutations in hematologic malignancies [54, 88–90]. We discovered splicing outliers of *CD79A* in samples MZL_3758 and LPL_0664, resulting in a truncation of 18 amino acids at the beginning of the 5th exon (Fig. 3E). AbSplice predicted that a rare 45-base-pair deletion spanning the acceptor site (NM_001783.4:c.568-2_610del) to be splice-affecting (AbSplice score = 0.36, Fig. 3F, Figure S17). This deletion, which was exclusively found in these two samples, is very likely the cause of this aberrant splicing event. Notably, similar deletions in the 4th and 5th exon of *CD79A* have been observed in diffuse large B-cell lymphoma, which has been shown to impact the ITAM signaling modules [88].

In summary, variants predicted to cause aberrant splicing by AbSplice and splicing outliers called by FRASER in transcriptomic data showed global enrichment for driver genes and proved invaluable in pinpointing affected driver genes in individual samples.

(See figure on next page.)

Fig. 3 Transcriptomic splicing outliers and genomic splice-affecting variants are enriched for hematologic malignancy driver genes. **A** Enrichment for CGC hematologic tumor suppressor genes among all genes called by FRASER and at most three significant genes per sample that were called aberrantly spliced from RNA-seq (FRASER). The genes are stratified by the number of samples in which a gene is called as a splicing outlier. Numbers of the genes and their nominal significance from the Fisher test are labeled at the top of the bars. Error bars of the odds ratio (95% confidence intervals) are shown where the Fisher's test is significant. **B** As in **A** among splice-affecting variants (AbSplice). **C** Overlap of splice-affecting variants, VEP splice-related variants, and splicing outliers on the sample-gene level. **D** As in **A** among splice-affecting variants with corresponding splicing outliers. **E** Junction counts (split reads) against total junction coverage (exon–intron or intron–exon spanning reads) of the displayed intron. The displayed intron of the *CD79A* only shows aberrant splicing in samples MZL_3758 and LPL_0664. **F** Sashimi plots showing RNA-seq read coverage (*y*-axis) and the numbers of split reads spanning an intron indicated on the exon-connecting line for two aberrant splicing events. Two case samples using a unique splice-site in *CD79A* 5th exon acceptor site and one control sample are displayed. The rare splice-affecting deletion (NM_001783.4:c.568-2_610del) predicted by AbSplice, which existed exclusively in the two case samples, is shown as black bars

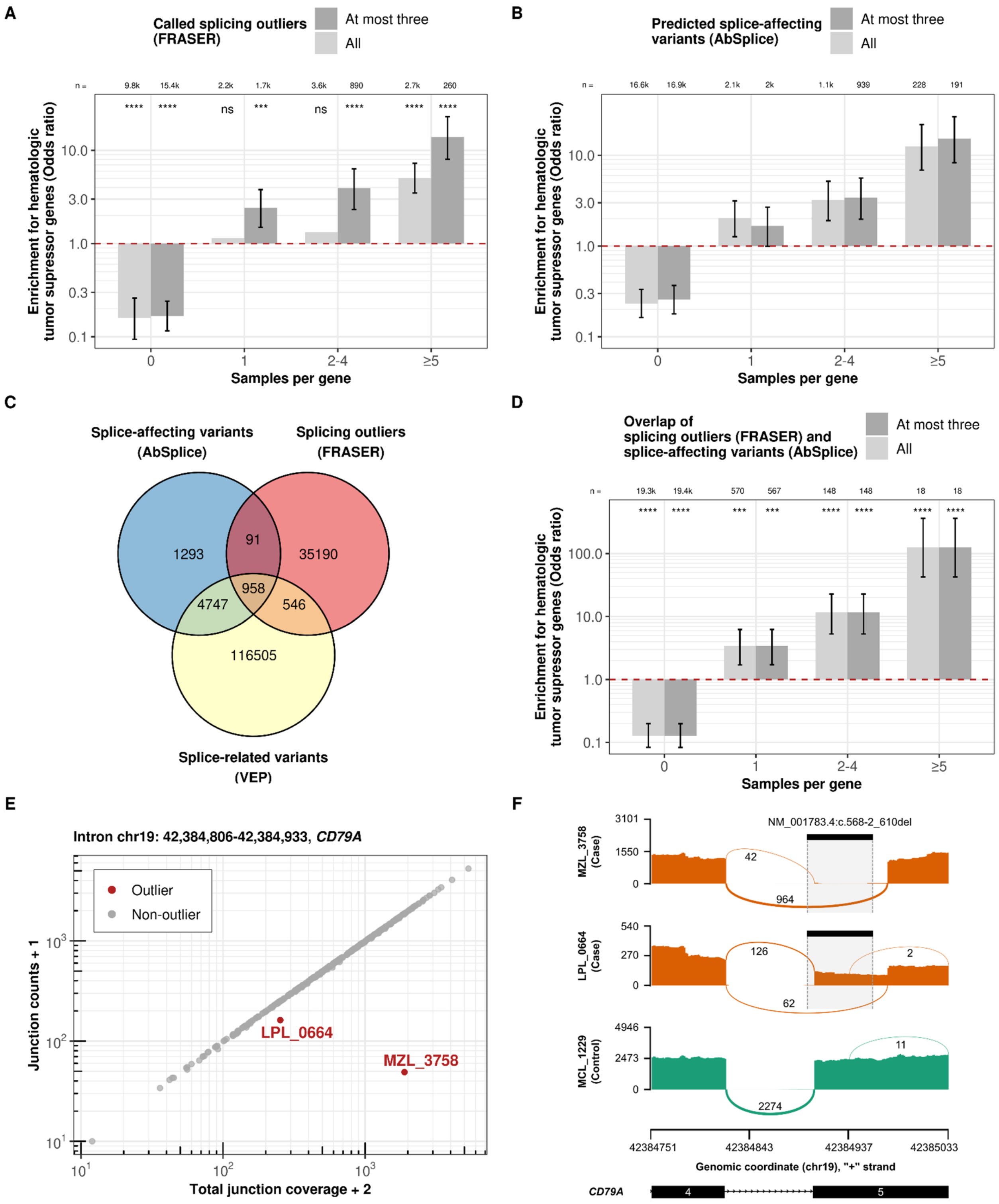


Fig. 3 (See legend on previous page.)

All our results from splice-affecting variants, expression outliers, splicing outliers, and recurrent mutational patterns aggregated by disease entities are provided (Availability of data and materials). This constitutes a comprehensive census detailing genomic and transcriptomic aberrations across 3760 hematologic malignancy samples.

Integration of rare genomic and transcriptomic aberrations improves hematologic malignancy driver gene prediction

We next trained models to predict driver genes based on genomic and transcriptomic features, including the seven IntOGen tools, AbSplice, OUTRIDER, NB-act, and FRASER. We used the complete dataset on the one hand and used the dataset stratified by 14 study groups on the other hand (Table 1). As ground truth, we considered the 322 CGC hematologic malignancy driver genes complemented with 55 additional curated hematologic panel genes (Methods). We further filtered the variants provided to the IntOGen pipeline based on variant allele frequency and sequencing depth, as this led to improved performance ("unmatched_filter_3", Methods, Figure S18). Moreover, we compared the performance of logistic regression, random forest, XGboost, and a fully connected neural network using fivefold cross-validation (Methods). We chose random forest as the final method as it showed the highest average precision based on the complete dataset (average precision 23%, Figure S19).

Using the complete dataset, we found that the genomic and transcriptomic features exhibited complementary predictive value for hematologic malignancy driver genes (Fig. 4A, B, Figure S20-S21). Specifically, integrating AbSplice variant effect predictions significantly enhanced the genomic-based model trained on the seven IntOGen tools. Moreover, we found that the transcriptomic features further significantly improved the model (Fig. 4B). These findings underscore the relevance of incorporating aberrant expression and splicing analyses to predict driver genes. We also explored models that include complementary features from external datasets. Specifically, we incorporated co-essentiality modules from DepMap [59] and a 256-dimensional functional gene embedding that integrates protein–protein interactions (PPI), genome-wide deletion screen results from the DepMap project, co-expression from bulk RNA-seq and single-cell RNA-seq compendia [60] (see Methods). This resulted in an enhancement of overall performance (average precision increased from 23 to 36%, Figure S22, Table S11 and S12). However, the inclusion of external data blurred the differences between disease entities and introduced predictions solely based on external gene functional evidence (Figure S23-S25). Given our focus on hematologic driver genes specifically, as opposed to cancer in general, and in order to highlight the added value of our dataset, we focussed on the model that does not incorporate external data in the following.

Among the 100 top-ranked genes, 63 were known from the ground truth, vastly greater than expectation (odds ratio = 106.3, $P = 1.6 \times 10^{-84}$, Fisher test; Fig. 4C, Table S13). These enrichments for well-known driver genes demonstrated the reliability of the model predictions. Four genes, *CDNKN1B*, *EIF3E*, *HLA-A*, and *IL6ST*, are CGC driver genes that have not been annotated as hematologic yet, indicating a broader role for those. Furthermore, 18 of these genes are targets of known, approved drugs (Fig. 4D, Table S14). Despite *TTN* being known to be false positive in mutational recurrence analysis as a long gene, the remaining genes were categorized as candidate drivers whose role remains to be further assessed.

In line with the results on the complete dataset, our integrative model outperformed IntOGen and MutSigCV—a driver gene predictor based on mutation frequency across all study groups (Figure S26). The study-group-wise models allowed further insights into disease-entity specificities (Fig. 4D, Table S15). Clustering the 100 top-ranked genes according to their study-group predicted probabilities revealed various disease entity specificities. Cluster 3 consisted of genes that exhibited high predicted probabilities across all disease entities. While *TP53* is a well-known pan-cancer tumor suppressor, we acknowledge that *ANKRD36*, *ANKRD36C*, and *UBC* may warrant further scrutiny as potential artifacts in global analysis [91]. Clusters 1 and 2 comprised genes that scored highly specifically in myeloid disease entities, such as *ASXL1* and *SRSF2* [9, 92]. Conversely, the genes in the remaining clusters were associated with specific disease entities. Notable examples include *NOTCH1*, *PHF6*, and *FBXW7*, which are well-acknowledged for their role in T-cell leukemia, as well as *BRAF*, which was exclusively predicted in the hairy-cell group, in accordance with its recognized function as a marker for HCL [10, 93–95].

Among the candidate genes, we identified several promising genes whose roles in hematologic malignancies are yet to be established. Our analysis predicted *SORL1* as a candidate driver in multiple study groups, including myelodysplastic neoplasm (precursor of AML), B-cell precursor ALL, and a study group comprising T-cell non-Hodgkin lymphoma and T-cell acute lymphoblastic leukemia. Consistent with these observations, *SORL1* was found to be expressed on the leukemic cell surface and released into plasma in AML and ALL, with its level decreasing during remission [96]. *EIF4G1* stood out as another interesting candidate driver in AML and lymphoid entities, supported by previous analyses suggesting that *EIF4G1* is involved in cell survival in AML as a downstream target of *MYCN*, a known oncogene in neuroblastoma [97]. We also found *TCP1* to be predicted in mastocytosis and B-cell precursor-ALL, whose high expression has been associated with AML drug resistance and poor survival through the activation of *AKT/mTOR* signaling [98]. A fourth exciting candidate, *ATXN1*, was predicted as a candidate driver in multiple disease entities. Its important paralog *ATXN1L* is known as a novel regulator of hematopoietic stem cell quiescence [99].

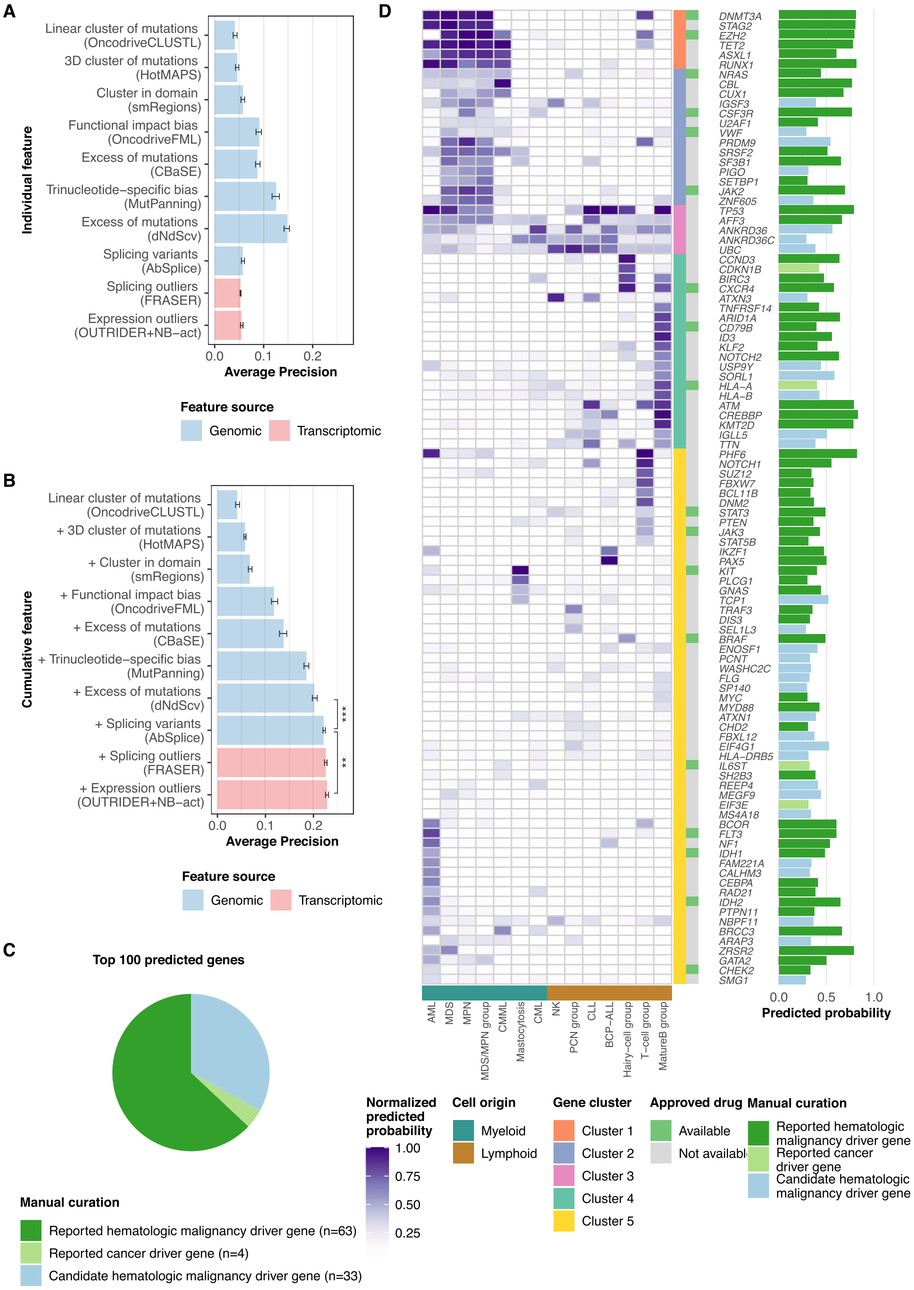


Fig. 4 Rare genomic and transcriptomic aberrations add value to hematologic malignancy driver gene prediction on top of methods detecting mutational recurrence. **A** Individual features and the corresponding model's performances (measured by average precision) using a random forest classifier. **B** Order as in **A** for cumulative features. The performance constantly improved when adding additional features. Asterisks denote nominal significance from the Wilcoxon test. **C** Numbers of genes in each category among the top 100 predicted genes when using all features and the complete dataset. Reported hematologic malignancy driver genes are the genes listed in either CGC hematologic malignancy driver genes or hematologic panel genes. Reported cancer driver genes are the genes documented by CGC. The rest of the genes are categorized as candidate hematologic malignancy driver genes. **D** The heatmap shows the predicted driver gene probability per gene (rows) and study group (columns) relative to the column-wise maximum value. Myeloid and lymphoid entities clustered due to the shared hematologic malignancy driver gene profiles (bottom track). The barplot shows the predicted probabilities from the model trained on the complete dataset. Bar colors as in **C**

Overall, our predictions based on 3760 tumor genomes and transcriptomes reveal promising hematologic malignancy driver candidates.

Our analysis is based on variants that include both somatic and rare germline variants, as these are hard to distinguish in the absence of matched control samples. A large fraction of germline variants is not necessarily a disadvantage for identifying driver genes, as they can encompass rare cancer-predisposing variants that would be discarded by a matched-control analysis. Moreover, somatic variants often include a large fraction of passenger variants that do not contribute to oncogenesis. Remarkably, the performances for calling CGC genes using each of the seven IntOGen tools in our dataset were similar to the performance reported by the IntOGen authors on TCGA for which matched control samples were available and thus allowed for trustable somatic variant calling (Figure S27, for disease entities with similar sample sizes). Although the disease entities differ, this analysis shows that our filtered variants are comparably predictive for cancer gene detection by cohort-wise gene-level enrichment approaches like IntOGen.

Outlier clustering identifies LRP1B as a potential marker in HCL-V and related B-cell malignancies

In addition to our global driver gene prediction analysis, we performed a detailed investigation into each disease entity, examining its association with expression outliers, splicing outliers, and variants predicted to cause aberrant splicing. Overall, we found 2716 significant associations between 11,273 genes and 24 disease entities (Table S16, Benjamini–Hochberg false discovery rate < 0.05, one-sided Fisher test). Focusing on activation outliers and annotated cancer driver genes, we found 43 associations between 37 CGC cancer driver genes and 12 disease entities (Fig. 5A). Some associations were already described in the literature (Table S17). For example, we confirmed that the transcription factors *TLX1* and *TLX3* were associated with T-cell acute lymphoblastic leukemia, in line with previous reports [100, 101]. In addition, other studies mentioned the role of *HOXA11*, *PREX2*, and *RET* in AML [102–104]. However, several associations have not yet been reported, including overexpression of *WNK2* in high-grade B-cell lymphoma, of *FAT4* in multiple myeloma, and of *LRP1B* in hairy cell leukemia (HCL), hairy cell leukemia variant (HCL-V), and marginal zone lymphoma (MZL).

Remarkably, *LRP1B* was associated with very rare entities. *LRP1B*, or Low-Density Lipoprotein Receptor-related protein 1B, is a frequently altered gene in multiple cancer types, but its exact role remains unclear [105]. In total, we found 24 samples (0.6% of all 3760 samples) with aberrantly high expressions of *LRP1B*. Thereof 21 were found within HCL-V, HCL, and MZL, where *LRP1B*-activated samples made up 44.8% (HCL-V), 7.4% (HCL), and 4.2% (MZL) of each entity (Fig. 5B). The other three cases with high *LRP1B* expression were found within multiple myeloma (MM), and no case was found in any of the other 20 entities. Among all *LRP1B*-activated samples, more than half of the cases (13 out of 24) were found in HCL-V patients, indicating that aberrant *LRP1B* expression might play an important role in HCL-V. While we observed a trend towards shorter overall survival (OS) of patients with *LRP1B* activation (Figure S28), this trend was not statistically significant, perhaps due to the low sample size (13 with *LRP1B* activation versus 16 without; median OS: 3.4 vs. 6.3 years; $P = 0.45$, two-sided log-rank test). The sample sizes were even lower for the other disease entities, hindering further statistical assessments (5 with *LRP1B* activation in HCL-V and 3 in MZL). We then evaluated *LRP1B* expression in an independent validation dataset of 42 samples. Consistent with observations made on our primary dataset, we detected 10 *LRP1B*-activated samples in HCL-V (5/14; 36%) and HCL (5/28; 18%, Fig. 5C) in the validation dataset. Moreover, the RNA-seq coverage for both datasets showed that samples overexpressing *LRP1B* expressed a truncated isoform in the majority (32/34; 94%) of samples (22/24 in the dataset; 10/10 in the validation dataset; Fig. 5D, Figure S29-S30). The two samples expressing full-length transcripts were both multiple myeloma, whose causes are yet to be understood. In the truncated isoform cases, *LRP1B* expression started at exon 13, thus lacking the first 636 amino acids of exon 1 to 12. Three start codons located within exon 13 could enable the start of translation using the canonical open reading frame. Assuming this transcript is translated, it yields a truncated LRP1B protein starting from the middle of the second β-propeller domain (Fig. 5E). However, we could not pinpoint a genomic cause for the truncated isoform, as no single-nucleotide variants, short insertions or deletions, structural variants, or gene fusions involving *LRP1B* were found to be specific to the affected samples.

Discussion

The study of rare cancer aberrations is an emerging field that greatly benefits from the growing large-scale next-generation sequencing [106, 107]. We have generated an extensive census of rare genomic and transcriptomic aberrations across 3760 patients spanning 24 hematologic malignancy disease entities. This census is based on the largest collection of hematologic malignancy samples that have undergone WGS along with matched RNA-seq data, which also includes rare disease entities. We cannot share without access restrictions the exact variants, expression outliers, and splicing outliers at the

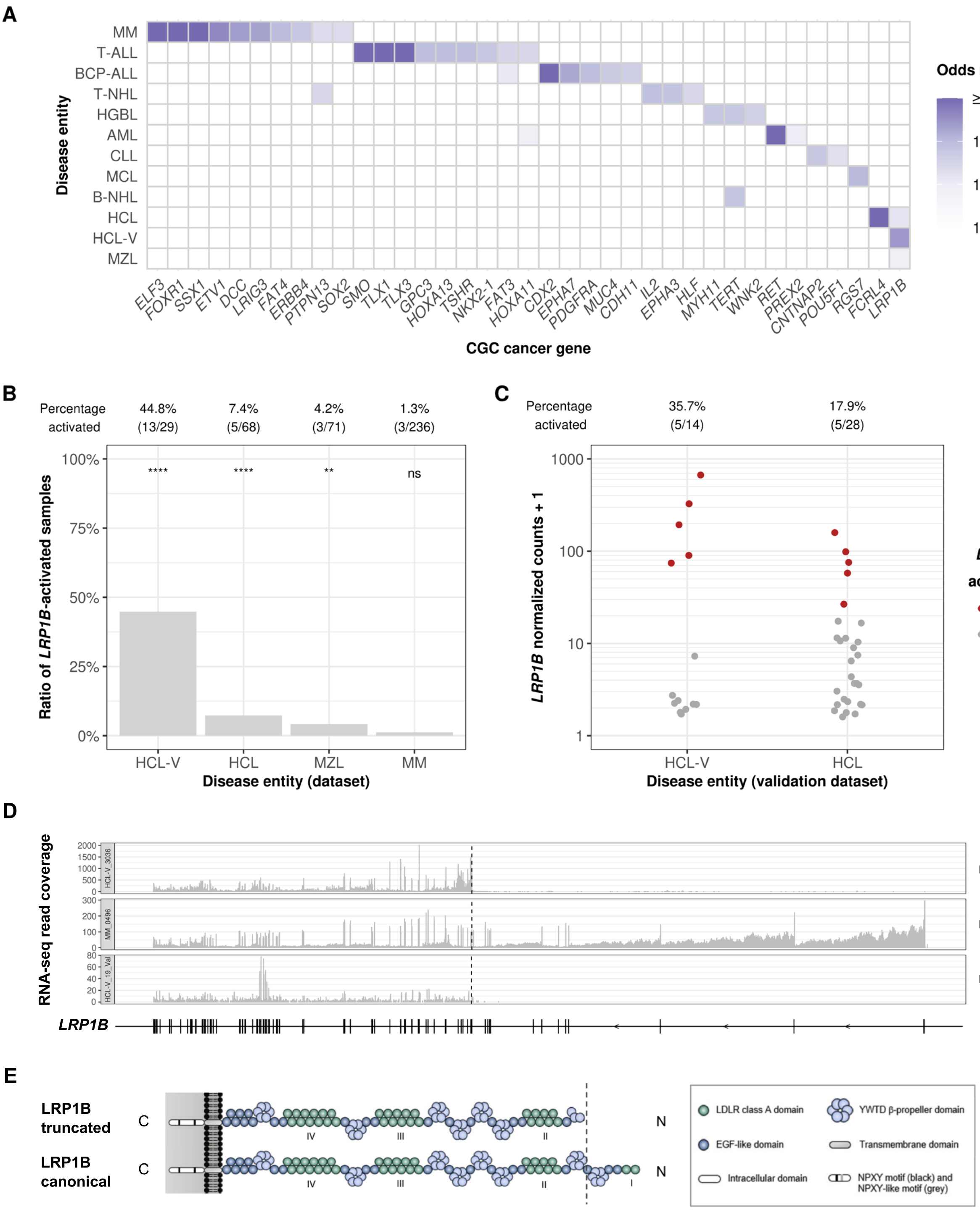


Fig. 5 Aberrant activation of *LRP1B* is predominant in HCL-V. **A** Disease entities (full names in Table 1) against aberrantly activated CGC genes, colored by odds ratio from Fisher test. **B** Percentage of *LRP1B*-activated samples in the four disease entities in which *LRP1B* activation occurred. Percentages of samples showing *LPR1B* activation (NB-act) and nominal significance from the one-sided Fisher test are labeled at the top. **C** Normalized counts of *LRP1B*-activated samples with regard to the different disease entities in the validation dataset. Percentages of samples showing *LPR1B* activation (Gaussian mixture clustering) are labeled at the top. **D** Transcriptomic coverage tracks showing different representative samples exhibiting LRP1B truncated isoforms (samples HCL-V_3036 dataset and HCL-V_19_Val validation dataset) and canonical isoform (sample MM_0496 dataset). For each track, the sample size annotation (n) refers to the number of LRP1B-activated samples with a similar RNA-seq coverage pattern. **E** Anticipated domain organization of truncated and canonical *LRP1B* (adapted from Príncipe et al. [105])

sample level, as such data would compromise research participant privacy. Nevertheless, we publicly provide our census (Availability of data and materials) and have demonstrated its utility. This census comprises frequencies of genes harboring variants predicted to cause aberrant splicing, expression outliers, and splicing outliers. All these categories were significantly enriched for known hematologic malignancy driver genes, highlighting their role as putative drivers in the corresponding samples. Notably, we have reaffirmed the well-established associations between transcription factors *TLX1* and *TLX3* with T-cell acute lymphoblastic leukemia, as well as *HOXA11*, *PREX2*, and *RET* with AML. Furthermore, we built pan-leukemia and entity-specific driver gene predictors by integrating this data, which successfully recovered known drivers and yielded promising novel candidates.

One of our notable findings is the identification of *LRP1B* as a potential marker for a subgroup of HCL-V and related B-cell malignancies. In our RNA-seq samples, *LRP1B* expression was rarely detected, occurring in only approximately 1% of cases. However, it was highly expressed in some cases of MZL, HCL, and approximately 50% of HCL-V, which we confirmed in an independent dataset. These three disease entities are mature B-cell malignancies and were previously regarded as separate entities in the revised 4th edition of the WHO classification of hematolymphoid tumors, distinguished based on immunophenotypic markers and molecular genetics [108]. HCL-V is typically resistant to conventional HCL therapy and does not show the HCL-specific *BRAF*-V600E mutation. However, as HCL, HCL-V, and MZL arise from malignant mature B-cells showing similar morphology, clear discrimination using conventional diagnostic techniques is often not possible. Thus, in the recently published 5th edition of WHO classification, the term “HCL-V” has been removed, recognizing that the biology of this disease is unrelated to HCL [8]. Instead, these cases are now considered splenic B-cell lymphoma/leukemia with prominent nucleoli (SBLPN), which also comprises rare cases of splenic MZL and B-prolymphocytic leukemia based on similar cytomorphological features. SBLPN rather serves as a placeholder for those morphologically defined cases of B-cell lymphoma not being classifiable into biologically distinct entities based on current evidence-based knowledge. We observed a tendency for a worse prognosis for HCL-V samples expressing *LRP1B*, though larger sample size is needed to establish a statistical significance. Overall, our results suggest a potential subcategorization of HCL-V/SBLPN based on *LRP1B* expression, whose functional implications remain to be elucidated.

LRP1B, also known as Low-Density Lipoprotein Receptor-related protein 1B, is broadly expressed in multiple normal tissues but not in blood or bone marrow [109–111]. It plays a role in various biological processes such as angiogenesis, chemotaxis, proliferation, adhesion, apoptosis, endocytosis, immunity, host-virus interaction, and protein folding. Additionally, *LRP1B* is among the most frequently altered genes in human cancer overall [112–117]. For tissues where *LRP1B* is normally expressed, *LRP1B* is often inactivated in cancer through several genetic and epigenetic mechanisms, making it a putative tumor suppressor gene. Overexpression of a truncated isoform of *LRP1B* in cells that normally do not express it, as we observed here, could play a role in oncogenesis by disrupting similar biological processes. Performing experiments on *LRP1B* is challenging. *LRP1B* is an extensive gene spanning 1.9 million bases, featuring 91 exons and a canonical transcript length of 16.5 kb. It is technically challenging to perform overexpression of the transcript of this enormous length [118]. Moreover, primary patient cells would be needed for downregulation experiments as *LRP1B* is not normally expressed in blood cells. Despite these challenges, our findings encourage further investigations to unravel the potential role of *LRP1B* in B-cell malignancies.

Our study has limitations: Our setting did not include matched control samples, preventing us from distinguishing somatic from rare germline variants. Although we found that working on this joint set of variants did not appear to hinder the performance of IntOGen for predicting cancer driver genes, we cannot delineate the contribution of somatic against rare germline variants to these cancers. Moreover, our approach is not geared towards retrieving common abnormalities and may underscore non-regulatory mutations. For example, *NPM1*, one of the most frequently mutated genes in AML [119], was not prioritized. *NPM1* was categorized as a driver gene by two IntOGen tools but not by the other five. Pathogenic mutations of *NPM1* act post-translationally by affecting the cellular localization of the NPM1 protein. Not surprisingly, the transcriptome data of *NPM1* were not informative. Consequently, *NPM1* was not prioritized by our driver gene prediction model. We have not characterized circulating DNA specifically, which has the potential to detect driver genes in hematologic malignancies [120] because its applicability varies within the different leukemia entities and is often limited due to its sensitivity [121]. Additionally, we did not include gene fusion and copy number variation calls, which are frequent causes of hematologic malignancies, into the driver gene prediction model, as we found them to have very high false positive rates during preliminary investigations. Moreover, we have purposely decided to restrict the input data of the driver gene prediction model to our dataset’s sole genomic and transcriptomic data to provide to the community predictions unbiased by previous literature. Future work could integrate this resource as prior information on pathways or protein interaction networks or with complementary datasets to provide a refined landscape of genomic and transcriptomic aberrations driving hematologic malignancies.

Conclusions

We established a unique and comprehensive census encompassing the genomic and transcriptomic landscape of 3760 hematologic malignancy samples, covering a wide range of disease entities. This comprehensive census can be leveraged to identify novel biomarkers, propose therapeutic decisions, and unravel the molecular underpinning of the heterogeneity of hematologic cancers.

Abbreviations

CGC	Cancer Gene Census
LRP1B	Low-Density Lipoprotein Receptor-related protein 1B
MLL	Munich Leukemia Laboratory
NB-act	Negative Binomial activation
RNA-Seq	Total RNA sequencing
VEP	Ensembl variant effect predictor
WGS	Whole genome sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-024-01331-6.

Additional file 1: Supplementary Figures, Supplementary Material and Methods, Supplementary Results. PDF file containing the accompanying supplementary text as well as all Supplementary Figures S1-S31.

Additional file 2: Table S1. Number of individuals, sex, and age groups aggregated by disease entities.

Additional file 3: Table S2. Number of filtered variants per sample.

Additional file 4: Table S3. Number of filtered variants aggregated by disease entities and genes.

Additional file 5: Table S4. Number of filtered variants aggregated by disease entities, genes and VEP consequences.

Additional file 6: Table S5. Number of underexpression outliers aggregated by disease entities and genes.

Additional file 7: Table S6. Number of overexpression outliers aggregated by disease entities and genes.

Additional file 8: Table S7. Number of activation outliers aggregated by disease entities and genes.

Additional file 9: Table S8. Number of splicing outliers aggregated by disease entities and genes.

Additional file 10: Table S9. Number of splice-affecting variants aggregated by disease entities and genes.

Additional file 11: Table S10. Fraction of splice-affecting variants within filtered variants aggregated by disease entities and genes.

Additional file 12: Table S11. Prediction result of the hematologic malignancy driver gene prediction model using the complete dataset and seven IntOGen tools, OUTRIDER, NB-act, FRASER, AbSplice, and external gene functional features.

Additional file 13: Table S12. Prediction result of the hematologic malignancy driver gene prediction model using each of the 14 study groups and seven IntOGen tools, OUTRIDER, NB-act, FRASER, AbSplice, and external gene functional features.

Additional file 14: Table S13. Prediction result of the hematologic malignancy driver gene prediction model using the complete dataset and seven IntOGen tools, OUTRIDER, NB-act, FRASER, AbSplice features.

Additional file 15: Table S14. Approved drugs of top-100 predicted genes using the complete dataset seven IntOGen tools, OUTRIDER, NB-act, FRASER, AbSplice.

Additional file 16: Table S15. Prediction result of the hematologic malignancy driver gene prediction model using each of the S14 study groups and seven IntOGen tools, OUTRIDER, NB-act, FRASER, AbSplice features.

Additional file 17: Table S16. Associations identified between disease entities and genes.

Additional file 18: Table S17. Curation of associations between CGC cancer driver genes disease entities using Activation outliers.

Additional file 19: Supplementary File 1. Mean copy ratio tracks aggregated by disease entities. Due to its large size, available at: Cao, Xueqi et al., Zenodo. https://zenodo.org/doi/10.5281/zenodo.8341456 (2024).

Additional file 20: Supplementary file 2. Mean FPKM matrix aggregated by disease entities and genes.

Additional file 21: Supplementary file 3. Fusion events aggregated by disease entities and genes.

Acknowledgements

The authors thank everyone from the MLL 5k project involved in patient inclusion, sample collection, and sequencing data generation; everyone involved in patient inclusion, sample collection, and sequencing data generation of the HCL-V validation dataset; Stefan Loipfinger, Luise Schuller, and Kim Anh Lilian Le for their contributions to the initial work of this project; Felix Brechtmann and Christian Mertes for the enlightening discussions. Parts of Fig. 1 were created with BioRender.com.

Authors' contributions

L.W., St.H., and J.G. designed the research; X.C., Sa.H., A.J.A., M.S., P.S., and St.H. performed the research; C.O. and St.H. collected the data; X.C., Sa.H., A.J.A., M.S., I.S., and N.W. analyzed and interpreted the data; X.C., Sa.H., and S.V. performed statistical analysis; X.C. and A.J.A. developed the software; M.H., L.W., St.H., and J.G. supervised the research; M.H., T.H., St.H., and J.G. acquired the funding; X.C., Sa.H., A.J.A., F.R.T., M.S., S.V., L.W., St.H., and J.G. wrote the manuscript. All authors read and approved the final manuscript.

Authors' information

Correspondence: Julien Gagneur, School of Computation, Information and Technology, Technical University of Munich, Boltzmannstraße 3, 85748 Garching bei München, Germany; e-mail: gagneur@in.tum.de; and Stephan Hutter, Münchner Leukämielabor GmbH (MLL), Max-Lebsche-Platz 31, 81377 München, Germany; e-mail: stephan.hutter@mll.com; and Leonhard Wachutka, School of Computation, Information and Technology, Technical University of Munich, Boltzmannstraße 3, 85748 Garching bei München, Germany; e-mail: wachutka@cit.tum.de.

Funding

Open Access funding enabled and organized by Projekt DEAL. This study was supported by the German Bundesministerium für Bildung und Forschung (BMBF) supported through the VALE (Entdeckung und Vorhersage der Wirkung von genetischen Varianten durch Artifizielle Intelligenz für LEukämie Diagnose und Subtypidentifizierung) project [031L0203B to X.Q., P.S., J.G.; 031L0203C to Sa.H., St.H.; and 031L0203A to S.V., M.H.]. This work was funded via the EVUK programme ("Next-generation AI for Integrated Diagnostics") of the Free State of Bavaria [to A.J.A. and J.G.].

Availability of data and materials

The raw genomic and transcriptomic data and the analyzed sample-level expression outliers, splicing outliers, and splice-affecting variants are not publicly available due to their potential to identify the individual and compromise research participants' privacy, but they are available from the corresponding author St.H. (stephan.hutter@mll.com) on reasonable request or by application via the Torsten Haferlach Research Foundation at https://torsten-haferlach-leukaemiediagnostik-stiftung.de/en/.

The aggregated gene-level expression outliers, splicing outliers, and splice-affecting variants, Supplementary Files and Supplementary Tables can be accessed at Cao, Xueqi et al., Aberrant transcriptome analysis of 3760 hematologic malignancies. Zenodo. https://zenodo.org/doi/10.5281/zenodo.8341456 (2024).

All code is available at: Cao, Xueqi et al., GitHub. https://github.com/gagneurlab/Leukemia_outlier (2024).

With the aggregated data from Zenodo, the Fig. 2ABC, Fig. 3ABD, Fig. 4, and Fig. 5AB can be recreated using: Cao, Xueqi et al., GitHub. https://github.com/gagneurlab/Leukemia_outlier/tree/main/driver_gene_prediction/Scripts/manuscript/agg_table_figure (2024).

An accompanying interactive website for data exploration and visualization is available at https://leukemia-outlier.cmm.cit.tum.de/

Declarations

Ethics approval and consent to participate

All patients or their legal guardians gave written informed consent for genetic analyses and to the use of laboratory results and clinical data for research purposes, according to the Declaration of Helsinki. The consent can be found at: https://www.mll.com/en/patient-consent-research-projects/mll-patient-consent-research-projects.pdf.

The study was approved by the MLL's institutional review board and by the Bavarian Ethics Committee, the Bavarian State Medical Association (Bayerische Landesärztekammer) with number 05117.

Consent for publication

Not applicable.

Competing interests

T.H. declares part ownership of Munich Leukemia Laboratory (MLL). Sa.H. and St.H. are employed by the MLL. The remaining authors declare that they have no competing interests.

Author details

[1]School of Computation, Information and Technology, Technical University of Munich, Garching, Germany. [2]Graduate School of Quantitative Biosciences (QBM), Munich, Germany. [3]Munich Leukemia Laboratory (MLL), Munich, Germany. [4]Institute of Biochemistry and Technical Biochemistry, University of Stuttgart, Stuttgart, Germany. [5]Department of Haematology, Oncology and Clinical Immunology, University Hospital Düsseldorf, Düsseldorf, Germany. [6]Division of Hematology, Department of Internal Medicine, The Ohio State University, Columbus, OH, USA. [7]Faculty of Biology, Ludwig-Maximilians-University Munich, Munich, Germany. [8]Computational Health Center, Helmholtz Center Munich, Neuherberg, Germany. [9]Helmholtz Association - Munich School for Data Science (MUDS), Munich, Germany. [10]Peter MacCallum Cancer Centre, Melbourne, Australia. [11]University of Melbourne, Melbourne, Australia. [12]Torsten Haferlach Leukämiediagnostik Stiftung, Munich, Germany. [13]Institute of Human Genetics, School of Medicine and Health, Technical University of Munich, Munich, Germany.

Received: 29 September 2023 Accepted: 4 April 2024
Published online: 20 May 2024

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Khoury JD, Solary E, Abla O, Akkari Y, Alaggio R, Apperley JF, et al. The 5th edition of the world health organization classification of Haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. Leukemia. 2022;36(7):1703–19.
3. de Matos SR, Shirasaki R, Downey-Kopyscinski SL, Matthews GM, Barwick BG, Gupta VA, et al. Genome-scale functional genomics identify genes preferentially essential for multiple myeloma cells compared to other neoplasias. Nat Cancer. 2023;4(5):754–73.
4. Ishio T, Kumar S, Shimono J, Daenthanasanmak A, Dubois S, Lin Y, et al. Genome-wide CRISPR screen identifies CDK6 as a therapeutic target in adult T-cell leukemia/lymphoma. Blood. 2022;139(10):1541–56.
5. Weber J, de la Rosa J, Grove CS, Schick M, Rad L, Baranov O, et al. PiggyBac transposon tools for recessive screening identify B-cell lymphoma drivers in mice. Nat Commun. 2019;10(1):1415.
6. Vantyghem S, Peterlin P, Thépot S, Ménard A, Dubruille V, Debord C, et al. Diagnosis and prognosis are supported by integrated assessment of next-generation sequencing in chronic myeloid malignancies. A real-life study. Haematologica. 2021;106(3):701–07.
7. Ramkissoon LA, Montgomery ND. Applications of next-generation sequencing in hematologic malignancies. Hum Immunol. 2021;82(11):859–70.
8. Alaggio R, Amador C, Anagnostopoulos I, Attygalle AD, de Araujo IBO, Berti E, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia. 2022;36(7):1720–48.
9. Arber DA, Orazi A, Hasserjian RP, Borowitz MJ, Calvo KR, Kvasnicka HM, et al. International consensus classification of myeloid neoplasms and acute leukemias: integrating morphologic, clinical, and genomic data. Blood. 2022;140(11):1200–28.
10. Campo E, Jaffe ES, Cook JR, Quintanilla-Martinez L, Swerdlow SH, Anderson KC, et al. The international consensus classification of mature lymphoid neoplasms: a report from the clinical advisory committee. Blood. 2022;140(11):1229–53.
11. Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: predictions and reality. Trends Mol Med. 2023;29(7):554–66.
12. Nussinov R, Tsai CJ, Jang H. Why are some driver mutations rare? Trends Pharmacol Sci. 2019;40(12):919–29.
13. Esai Selvan M, Onel K, Gnjatic S, Klein RJ, Gümüş ZH. Germline rare deleterious variant load alters cancer risk, age of onset and tumor characteristics. Npj Precis Oncol. 2023;7(1):1–12.
14. Kuan-lin Huang, R. Jay Mash, Yige Wu et al. Cell. 173;(2)355–70.e14. https://doi.org/10.1016/j.cell.2018.03.039.
15. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MDM, Wendl MC, et al. Integrated analysis of germline and somatic variants in ovarian cancer. Nat Commun. 2014;5(1):3156.
16. Lu C, Xie M, Wendl MC, Wang J, McLellan MD, Leiserson MDM, et al. Patterns and functional implications of rare germline variants across 12 cancer types. Nat Commun. 2015;6(1):10086.
17. Orgueira AM, López CM, Raíndo PA, Arias JÁD, Rodríguez AB, Pérez LB, et al. Detection of rare germline variants in the genomes of patients with B-cell neoplasms. Cancers. 2021;13(6):1340.
18. Vali-Pour M, Park S, Espinosa-Carrasco J, Ortiz-Martínez D, Lehner B, Supek F. The impact of rare germline variants on human somatic mutation processes. Nat Commun. 2022;13(1):3724.
19. Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, Kahles A, et al. Genomic basis for RNA alterations in cancer. Nature. 2020;578(7793):129–36.
20. Hautin M, Mornet C, Chauveau A, Bernard D, Corcos L, Lippert E. Splicing Anomalies in Myeloproliferative Neoplasms: Paving the Way for New Therapeutic Venues. Cancers. 2020;12(8):2216.
21. Grinev VV, Barneh F, Ilyushonak IM, Nakjang S, Smink J, Van Oort A, et al. RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia. Nat Commun. 2021;12(1):520.
22. Tanaka A, Nakano TA, Nomura M, Yamazaki H, Bewersdorf JP, Mulet-Lazaro R, et al. Aberrant EVI1 splicing contributes to EVI1-rearranged leukemia. Blood. 2022;140(8):875–88.
23. Obeng EA. Mutant SF3B1 splices a more leukemogenic EVI1. Blood. 2022;140(8):800–1.
24. Huber S, Haferlach T, Meggendorfer M, Hutter S, Hoermann G, Baer C, et al. SF3B1 mutations in AML are strongly associated with MECOM rearrangements and may be indicative of an MDS pre-phase. Leukemia. 2022;36(12):2927–30.

25. Yang YT, Chiu YC, Kao CJ, Hou HA, Lin CC, Tsai CH, et al. The prognostic significance of global aberrant alternative splicing in patients with myelodysplastic syndrome. Blood Cancer J. 2018;8(8):78.
26. Szelest M, Masternak M, Zając M, Chojnacki M, Skórka K, Zaleska J, et al. The role of NPM1 alternative splicing in patients with chronic lymphocytic leukemia. PLoS One. 2022;17(10):e0276674.
27. Dlamini Z, Shoba B, Hull R. Splicing machinery genomics events in acute myeloid leukaemia (AML): in search for therapeutic targets, diagnostic and prognostic biomarkers. Am J Cancer Res. 2020;10(9):2690–704.
28. Rivera OD, Mallory MJ, Quesnel-Vallières M, Chatrikhi R, Schultz DC, Carroll M, et al. Alternative splicing redefines landscape of commonly mutated genes in acute myeloid leukemia. Proc Natl Acad Sci. 2021;118(15):e2014967118.
29. Black KL, Naqvi AS, Asnani M, Hayer KE, Yang SY, Gillespie E, et al. Aberrant splicing in B-cell acute lymphoblastic leukemia. Nucleic Acids Res. 2018;46(21):11357–69.
30. Huber S, Haferlach T, Meggendorfer M, Hutter S, Hoermann G, Summerer I, et al. Mutations in spliceosome genes in myelodysplastic neoplasms and their association to ring sideroblasts. Leukemia. 2023;37(2):500–2.
31. Lewis RE, Cruse JM, Sanders CM, Webb RN, Suggs JL. Aberrant expression of T-cell markers in acute myeloid leukemia. Exp Mol Pathol. 2007;83(3):462–3.
32. Collins S, Coleman H, Groudine M. Expression of bcr and bcr-abl fusion transcripts in normal and leukemic cells. Mol Cell Biol. 1987;7(8):2870–6.
33. Rimokh R, Berger F, Bastard C, Klein B, French M, Archimbaud E, et al. Rearrangement of CCND1 (BCL1/PRAD1) 3′ untranslated region in mantle- cell lymphomas and t(11q13)-associated leukemias. Blood. 1994;83(12):3689–96.
34. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 2016;35(19):2413–27.
35. Nakai H, Kaneko H, Horiike S, Ariyama Y, Misawa S, Kashima K, et al. Multiple aberrant splicing of the p53 transcript without genomic mutations around exon-intron junctions in a case of chronic myelogenous leukaemia in blast crisis: a possible novel mechanism of p53 inactivation. Br J Haematol. 1994;87(4):839–42.
36. Wagner N, Çelik MH, Hölzlwimmer FR, Mertes C, Prokisch H, Yépez VA, et al. Aberrant splicing prediction across human tissues. Nat Genet. 2023;55(5):861–70.
37. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555–72.
38. Brechtmann F, Mertes C, Matusevičiūtė A, Yépez VA, Avsec Ž, Herzog M, et al. OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. Am J Hum Genet. 2018;103(6):907–17.
39. 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;110(12):2056–67.
40. Kern W, Voskova D, Schoch C, Hiddemann W, Schnittger S, Haferlach T. Determination of relapse risk based on assessment of minimal residual disease during complete remission by multiparameter flow cytometry in unselected patients with acute myeloid leukemia. Blood. 2004;104(10):3078–85.
41. Haferlach T, Schoch C, Löffler H, Gassmann W, Kern W, Schnittger S, et al. Morphologic Dysplasia in De Novo Acute Myeloid Leukemia (AML) is related to unfavorable cytogenetics but has no independent prognostic relevance under the conditions of intensive induction therapy: results of a multiparameter analysis from the German AML Cooperative Group Studies. J Clin Oncol. 2003;21(2):256–65.
42. Schoch C, Schnittger S, Bursch S, Gerstner D, Hochhaus A, Berger U, et al. Comparison of chromosome banding analysis, interphase- and hypermetaphase-FISH, qualitative and quantitative PCR for diagnosis and for follow-up in chronic myeloid leukemia: a study on 350 cases. Leukemia. 2002;16(1):53–9.
43. Haferlach C, Walter W, Stengel A, Meggendorfer M, Hutter S, Kern W, et al. The diverse landscape of fusion transcripts in 25 different hematological entities. Leuk Lymphoma. 2021;62(13):3292–5.
44. Hershberger CE, Moyer DC, Adema V, Kerr CM, Walter W, Hutter S, et al. Complex landscape of alternative splicing in myeloid neoplasms. Leukemia. 2021;35(4):1108–20.
45. Stengel A, Baer C, Walter W, Meggendorfer M, Kern W, Haferlach T, et al. Mutational patterns and their correlation to CHIP-related mutations and age in hematological malignancies. Blood Adv. 2021;5(21):4426–34.
46. Höllein A, Twardziok SO, Walter W, Hutter S, Baer C, Hernandez-Sanchez JM, et al. The combination of WGS and RNA-Seq is superior to conventional diagnostic tests in multiple myeloma: Ready for prime time? Cancer Genet. 2020;242:15–24.
47. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–73.
48. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
49. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013;41(6):e67.
50. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020;581(7809):444–51.
51. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.
52. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
53. Yépez VA, Mertes C, Müller MF, Klaproth-Andrade D, Wachutka L, Frésard L, et al. Detection of aberrant gene expression events in RNA sequencing data. Nat Protoc. 2021;16(2):1276–96.
54. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696–705.
55. Auwera G., O'Connor B. Genomics in the Cloud. O'Reilly. 2020.
56. Therneau TM. A Package for Survival Analysis in R; 2023. https://CRAN.R-project.org/package=survival.
57. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using "ggplot2". 2021. https://CRAN.Rproject.org/package=survminer.
58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12(85):2825–30.
59. Wainberg M, Kamber RA, Balsubramani A, Meyers RM, Sinnott-Armstrong N, Hornburg D, et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat Genet. 2021;53(5):638–49.
60. Brechtmann F, Bechtler T, Londhe S, Mertes C, Gagneur J. Evaluation of input data modality choices on functional gene embeddings. NAR Genom Bioinform. 2023;5(4):lqad095.
61. Kelleher KJ, Sheils TK, Mathias SL, Yang JJ, Metzger VT, Siramshetty VB, et al. Pharos 2023: an integrated resource for the understudied human proteome. Nucleic Acids Res. 2023;51(D1):D1405–16.
62. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
63. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
64. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. 2016;8(1):289–317.
65. Arnedo-Pac C, Mularoni L, Muiños F, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. Bioinformatics. 2019;35(22):4788–90.
66. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. 2016;17(1):128.
67. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res. 2016;76(13):3719–31.
68. Martínez-Jiménez F, Muiños F, López-Arribillaga E, Lopez-Bigas N, Gonzalez-Perez A. Systematic analysis of alterations in the ubiquitin

proteolysis system reveals its contribution to driver mutations in cancer. Nat Cancer. 2020;1(1):122–35.

69. Weghorn D, Sunyaev S. Bayesian inference of negative and positive selection in human cancers. Nat Genet. 2017;49(12):1785–8.
70. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, et al. Identification of cancer driver genes based on nucleotide context. Nat Genet. 2020;52(2):208–18.
71. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal patterns of selection in cancer and somatic tissues. Cell. 2017;171(5):1029-1041.e21.
72. Dekker J, Schot R, Bongaerts M, de Valk WG, van Veghel-Plandsoen MM, Monfils K, et al. Web-accessible application for identifying pathogenic transcripts with RNA-seq: increased sensitivity in diagnosis of neurodevelopmental disorders. Am J Hum Genet. 2023;110(2):251–72.
73. Lunke S, Bouffler SE, Patel CV, Sandaradura SA, Wilson M, Pinner J, et al. Integrated multi-omics for rapid rare disease diagnosis on a national scale. Nat Med. 2023;29(7):1681–91.
74. Lee M, Kwong AKY, Chui MMC, Chau JFT, Mak CCY, Au SLK, et al. Diagnostic potential of the amniotic fluid cells transcriptome in deciphering mendelian disease: a proof-of-concept. NPJ Genomic Med. 2022;28(7):74.
75. Vialle RA, de Paiva LK, Bennett DA, Crary JF, Raj T. Integrating whole-genome sequencing with multi-omic data reveals the impact of structural variants on gene regulation in the human brain. Nat Neurosci. 2022;25(4):504–14.
76. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013;155(4):948–62.
77. Watkins TBK, Lim EL, Petkovic M, Elizalde S, Birkbak NJ, Wilson GA, et al. Pervasive chromosomal instability and karyotype order in tumour evolution. Nature. 2020;587(7832):126–32.
78. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, et al. The impact of rare variation on gene expression across tissues. Nature. 2017;550(7675):239–43.
79. Jiang S. Tet2 at the interface between cancer and immunity. Commun Biol. 2020;3(1):667.
80. Weissmann S, Alpermann T, Grossmann V, Kowarsch A, Nadarajah N, Eder C, et al. Landscape of TET2 mutations in acute myeloid leukemia. Leukemia. 2012;26(5):934–42.
81. Bensberg M, Rundquist O et al. TET2 as a tumor suppressor and therapeutic target in T-cell acute lymphoblastic leukemia. Proc Natl Acad Sci. 2021;118(34):e2110758118.
82. Feng Y, Li X, Cassady K, Zou Z, Zhang X. TET2 Function in Hematopoietic Malignancies, Immune Regulation, and DNA Repair. Front Oncol. 2019;9:210.
83. Zeisig BB, Kulasekararaj AG, Mufti GJ, Eric So CW. SnapShot: acute myeloid leukemia. Cancer Cell. 2012;22(5):698-698.e1.
84. Di Fiore R, D'Anneo A, Tesoriere G, Vento R. RB1 in cancer: different mechanisms of RB1 inactivation and alterations of pRb pathway in tumorigenesis. J Cell Physiol. 2013;228(8):1676–87.
85. Yépez VA, Gusic M, Kopajtich R, Mertes C, Smith NH, Alston CL, et al. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med. 2022;14(1):38.
86. Cheng J, Nguyen TYD, Cygan KJ, Çelik MH, Fairbrother WG, Avsec Ž, et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. 2019;20(1):48.
87. Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535–548.e24.
88. Davis RE, Ngo VN, Lenz G, Tolar P, Young RM, Romesser PB, et al. Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. Nature. 2010;463(7277):88–92.
89. Sakatani A, Igawa T, Okatani T, Fujihara M, Asaoku H, Sato Y, et al. Clinicopathological significance of CD79a expression in classic Hodgkin lymphoma. J Clin Exp Hematop. 2020;60(3):78–86.
90. Lenk L, Carlet M, Vogiatzi F, Spory L, Winterberg D, Cousins A, et al. CD79a promotes CNS-infiltration and leukemia engraftment in pediatric B-cell precursor acute lymphoblastic leukemia. Commun Biol. 2021;4(1):1–6.
91. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol. 2010;2(1):a001008.
92. Lindsley RC, Mar BG, Mazzola E, Grauman PV, Shareef S, Allen SL, et al. Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. Blood. 2015;125(9):1367–76.
93. Falini B, Martelli MP, Tiacci E. BRAF V600E mutation in hairy cell leukemia: from bench to bedside. Blood. 2016;128(15):1918–27.
94. Mullighan CG. Mutations of NOTCH1, FBXW7, and prognosis in T-lineage acute lymphoblastic leukemia. Haematologica. 2009;94(10):1338–40.
95. Van Vlierberghe P, Palomero T, Khiabanian H, Van der Meulen J, Castillo M, Van Roy N, et al. PHF6 mutations in T-cell acute lymphoblastic leukemia. Nat Genet. 2010;42(4):338–42.
96. Sakai S, Nakaseko C, Takeuchi M, Ohwada C, Shimizu N, Tsukamoto S, et al. Circulating soluble LR11/SorLA levels are highly increased and ameliorated by chemotherapy in acute leukemias. Clin Chim Acta. 2012;413(19):1542–8.
97. Peramangalam PS, Surapally S, Zheng S, Burns R, Zhu N, Rao S, et al. MYCN regulates cell survival via EIF4G1 in Inv(16) acute myeloid leukemia. Blood. 2022;140(Supplement 1):9117–8.
98. Chen X, Chen X, Huang Y, Lin J, Wu Y, Chen Y. TCP1 increases drug resistance in acute myeloid leukemia by suppressing autophagy via activating AKT/mTOR signaling. Cell Death Dis. 2021;12(11):1058.
99. Kahle JJ, Souroullas GP, Yu P, Zohren F, Lee Y, Shaw CA, et al. Ataxin1L Is a regulator of HSC function highlighting the utility of cross-tissue comparisons for gene discovery. PLoS Genet. 2013;9(3):e1003359.
100. Girardi T, Vicente C, Cools J, De Keersmaecker K. The genetics and molecular biology of T-ALL. Blood. 2017;129(9):1113–23.
101. Riz I, Hawley TS, Johnston H, Hawley RG. Role of TLX1 in T-cell acute lymphoblastic leukemia pathogenesis. Br J Haematol. 2009;145(1):140–3.
102. Rudat S, Pfaus A, Bühler C, Rabe S, Bullinger L, Gröschel S, et al. The RET receptor tyrosine kinase promotes acute myeloid leukemia through protection of FLT3-ITD mutants from autophagic degradation. Blood. 2016;128(22):2849.
103. McNeer NA, Philip J, Geiger H, Ries RE, Lavallée VP, Walsh M, et al. Genetic mechanisms of primary chemotherapy resistance in pediatric acute myeloid leukemia. Leukemia. 2019;33(8):1934–43.
104. Fu JF, Shih LY, Yen TH. HOXA11 plays critical roles in disease progression and response to cytarabine in AML. Oncol Rep. 2021;46(1):150.
105. Príncipe C, de DionísioSousa IJ, Prazeres H, Soares P, Lima RT. LRP1B: a giant lost in cancer translation. Pharmaceuticals. 2021;14(9):836.
106. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20.
107. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371-385.e18.
108. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, Thiele J. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. 4th Edition, Volume 2. IARC. 2017.
109. Strickland DK, Gonias SL, Argraves WS. Diverse roles for the LDL receptor family. Trends Endocrinol Metab. 2002;13(2):66–74.
110. May P, Woldt E, Matz RL, Boucher P. The LDL receptor-related protein (LRP) family: an old family of proteins with new physiological functions. Ann Med. 2007;39(3):219–28.
111. Dieckmann M, Dietrich MF, Herz J. Lipoprotein receptors – an evolutionarily ancient multifunctional receptor family. Biol Chem. 2010;391(11):1341–63.
112. Beer AG, Zenzmaier C, Schreinlechner M, Haas J, Dietrich MF, Herz J, et al. Expression of a recombinant full-length LRP1B receptor in human non-small cell lung cancer cells confirms the postulated growth-suppressing function of this large LDL receptor family member. Oncotarget. 2016;7(42):68721–33.
113. Haas J, Beer AG, Widschwendter P, Oberdanner J, Salzmann K, Sarg B, et al. LRP1b shows restricted expression in human tissues and binds to several extracellular ligands, including fibrinogen and apoE – carrying lipoproteins. Atherosclerosis. 2011;216(2):342–7.
114. Liu CX, Ranganathan S, Robinson S, Strickland DK. γ-Secretase-mediated Release of the Low Density Lipoprotein Receptor-related

Protein 1B Intracellular Domain Suppresses Anchorage-independent Growth of Neuroglioma Cells. J Biol Chem. 2007;282(10):7504–11.
115. Brown LC, Tucker MD, Sedhom R, Schwartz EB, Zhu J, Kao C, et al. LRP1B mutations are associated with favorable outcomes to immune checkpoint inhibitors across multiple cancer types. J Immunother Cancer. 2021;9(3):e001792.
116. Cowin PA, George J, Fereday S, Loehrer E, Van Loo P, Cullinane C, et al. LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin. Cancer Res. 2012;72(16):4060–73.
117. Liu F, Hou W, Liang J, Zhu L, Luo C. LRP1B mutation: a novel independent prognostic factor and a predictive tumor mutation burden in hepatocellular carcinoma. J Cancer. 2021;12(13):4039–48.
118. Príncipe C, de Sousa IJD, Prazeres H, Soares P, Lima RT. LRP1B: a giant lost in cancer translation. Pharmaceuticals. 2021;14(9):836.
119. Falini B, Brunetti L, Sportoletti P, Martelli MP. NPM1-mutated acute myeloid leukemia: from bench to bedside. Blood. 2020;136(15):1707–21.
120. Ogawa M, Yokoyama K, Imoto S, Tojo A. Role of circulating tumor DNA in hematological malignancy. Cancers. 2021;13(9):2078.
121. Talotta D, Almasri M, Cosentino C, Gaidano G, Moia R. Liquid biopsy in hematological malignancies: current and future applications. Front Oncol. 2023;13:1164517.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.