Clinical Chemistry 66:12
1521–1530 (2020)

Targeted RNA-Sequencing Enables Detection of Relevant Translocations and Single Nucleotide Variants and Provides a Method for Classification of Hematological Malignancies–RANKING

Kim de Lange,[a,b,†] Eddy N. de Boer,[a,†] Anneke Bosga,[a] Mohamed Z. Alimohamed,[a,c] Lennart F. Johansson,[a,b] André B. Mulder,[d] Edo Vellenga,[e] Cleo C. van Diemen,[a] Patrick Deelen,[a,b,f] Eva van den Berg,[a,†] and Birgit Sikkema-Raddatz[a,*]

BACKGROUND: Patients with hematological malignancies (HMs) carry a wide range of chromosomal and molecular abnormalities that impact their prognosis and treatment. Since no current technique can detect all relevant abnormalities, technique(s) are chosen depending on the reason for referral, and abnormalities can be missed. We tested targeted transcriptome sequencing as a single platform to detect all relevant abnormalities and compared it to current techniques.

MATERIAL AND METHODS: We performed RNA-sequencing of 1385 genes (TruSight RNA Pan-Cancer, Illumina) in bone marrow from 136 patients with a primary diagnosis of HM. We then applied machine learning to expression profile data to perform leukemia classification, a method we named RANKING. Gene fusions for all the genes in the panel were detected, and overexpression of the genes *EVI1*, *CCND1*, and *BCL2* was quantified. Single nucleotide variants/indels were analyzed in acute myeloid leukemia (AML), myelodysplastic syndrome and patients with acute lymphoblastic leukemia (ALL) using a virtual myeloid (54 genes) or lymphoid panel (72 genes).

RESULTS: RANKING correctly predicted the leukemia classification of all AML and ALL samples and improved classification in 3 patients. Compared to current methods, only one variant was missed, c.2447A>T in KIT (RT-PCR at 10^{-4}), and BCL2 overexpression was not seen due to a t(14; 18)(q32; q21) in 2% of the cells. Our RNA-sequencing method also identified 6 additional fusion genes and overexpression of CCND1 due to a t(11; 14)(q13; q32) in 2 samples.

CONCLUSIONS: Our combination of targeted RNA-sequencing and data analysis workflow can improve the detection of relevant variants, and expression patterns can assist in establishing HM classification.

Introduction

The classification, progression prognosis, and treatment of hematological malignancies (HMs) is based on the identification of specific genetic and molecular abnormalities. Rapid and comprehensive genetic diagnosis is therefore essential. HMs are currently classified based on recurrent cytogenetic abnormalities and gene mutations (1) such as translocations, copy number variants (CNVs), and single nucleotide variants (SNVs). A broad range of labor-intensive and expensive techniques are used to detect these abnormalities, including karyotyping, fluorescence in situ hybridization (FISH), reverse transcription (RT)-PCR, array, and targeted next-generation sequencing (NGS). Karyotyping is used to detect numeric and recurrent chromosomal aberrations, but this method has only limited genomic resolution and requires mitotic cells. FISH and RT-PCR are less effective in detecting translocations in genes with multiple possible breakpoints and gene partners (2). The development of single nucleotide polymorphism (SNP)-array-based platforms using DNA can now provide higher levels of resolution in comparison to karyotyping and do not require cell culturing. These arrays fall short, however, in the detection of balanced translocations,

[a]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; [b]Genomics Coordination Center, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands; [c]Department of Haematology and Blood Transfusion, Muhimbili University of Health and Allied Sciences, Dar-es-Salaam, Tanzania; [d]Department of Laboratory Medicine, University of Groningen, University Medical Center Groningen, the Netherlands; [e]Department of Hematology, University of Groningen, University Medical Center Groningen, the Netherlands; [f]Department of Genetics, University Medical Centre Utrecht, Utrecht, the Netherlands.

*Address correspondence to this author at: Department of Genetics, University Medical Center Groningen P.O. Box 30001, 9700 RB Groningen, The Netherlands. E-mail b.sikkema01@umcg.nl.

[†]Contributed equally.

Received May 1, 2020; accepted August 31, 2020.
DOI: 10.1093/clinchem/hvaa221

which are the model processes behind gene fusions (3). However, arrays do detect CNVs to determine complex aberrations and hypo- and hyperdiploidy. Arrays using RNA as input have also been used to predict cancer classes, i.e., to distinguish between the acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) types of leukemia based on expression monitoring (4). In addition, targeted NGS techniques have been successfully applied to identify SNVs, indels, and translocations (5).

Because no one technique can detect all the genetic abnormalities relevant in HM, diagnosticians have to make choices about which combination of techniques should be used and which genes should be investigated. These choices are made based on the referral type of HM, and relevant abnormalities may go undetected where a diagnosis could have been made by choosing another technique or investigating other genes. RNA-sequencing has the potential to overcome this in one technique, as has been demonstrated for the AML type of leukemia (6). While whole transcriptome sequencing enables the analysis of all genes in one test, targeted approaches allow deeper sequencing for the same amount of sample. Reaching the required sensitivity is particularly important for HM samples because the percentage of tumor cells in bone marrow can be low.

We implemented targeted RNA-sequencing as a single platform to acquire all the genetic information relevant for HM classification and risk categorization. Our aim was to develop a single workflow to detect SNVs/indels, gene fusions, and overexpression of relevant genes. CNV analysis and tandem duplication detection were not part of this study. We used machine learning applied to expression profiling to predict the type of HM, a method we named RANKING (expRession profiling for clAssificatioN of leuKemias using modeling). RANKING allows users to choose which data out of the transcriptome are needed for analysis. We then applied our workflow to samples from patients with a primary diagnosis of HM and compared the outcome of our method with that of current genetic tests.

Materials and Methods

PATIENT BONE MARROW CELLS

Bone marrow cells were obtained from 136 patients with a primary diagnosis of HM, excluding chronic lymphocytic leukemia, and multiple myeloma, following informed consent. Samples were included from November 2016–October 2017. The study protocol was approved by the Ethics Committee of the University Medical Centre Groningen (METC 2014.051, 10-2-2014).

Mononuclear cells were isolated on the same day as bone marrow withdrawal using lymphoprep (Lymphoprep^{TM}), following the manufacturer's protocol, and stored in RNAprotect (Qiagen, catalog no. 76526, Thermo Fisher Scientific) at −20 °C. Cells were homogenized using QIAshredder (Qiagen, catalog no. 79654), followed by RNA isolation using the RNAeasy Plus Mini Kit (Qiagen, catalog no. 74104) according to the manufacturer's protocol. RNA quality was measured using Fragment Analyzer (Applied Biosystems, Thermo Fisher Scientific), and samples with an RNA Quality Number >7 and electrograms with both 18 and 28 s peaks were selected.

STUDY DESIGN: VARIANT DETECTION FROM TARGETED RNA-SEQUENCING TESTING

The workflow for the data analysis of the targeted RNA-sequencing is outlined in Fig. 1. In general, we compared the clinically relevant abnormalities reported by routine genetic tests to the outcome of our analysis using targeted RNA-sequencing (number of currents tests in Fig. 1 in the online Data Supplement). We performed SNV/indel detection and fusion gene detection and measured overexpression to detect different types of abnormalities. For each type of analysis, the sample cohort was split into a validation and an application set (Table 1).

For SNV/indel analysis, based on the HM classification for which the patient was referred, we used either a virtual myeloid [AML and myelodysplastic syndrome (MDS), 54 genes] or lymphoid (ALL, 72 genes) panel to filter relevant variants in these genes out of the RNA-sequencing data (see Supplemental Table 1). In total, 90 samples out of the 136 matched the classifications AML, MDS, and ALL. For validation, 63 samples were available where variants were previously detected using RT-PCR analysis or Sanger sequencing for the genes *NPM1*, *CEBPA*, and *KIT* ($N = 35$) and with the Illumina TrueSight Myeloid sequencing panel ($N = 28$). We analyzed an application set of 27 samples ($N = 13$ MDS and $N = 14$ ALL) in which no SNVs/indels had been detected before, and the variants identified here were confirmed by Sanger sequencing, Multiplex-dependent ligation probe amplification (MLPA), or digital droplet (dd) PCR in DNA and/or RNA of the same material.

For fusion gene detection, the first 40 samples collected of the 136 samples were used as a validation set to determine thresholds for data interpretation. An application set of 96 samples was then analyzed blind using the criteria determined using the first 40 samples.

For overexpression detection, we used the 84 samples of the 136 where we were convinced about the HM classification. We first performed a principle component analysis (PCA) (7) on our expression data. Using the first 5 principle components, which explained most of

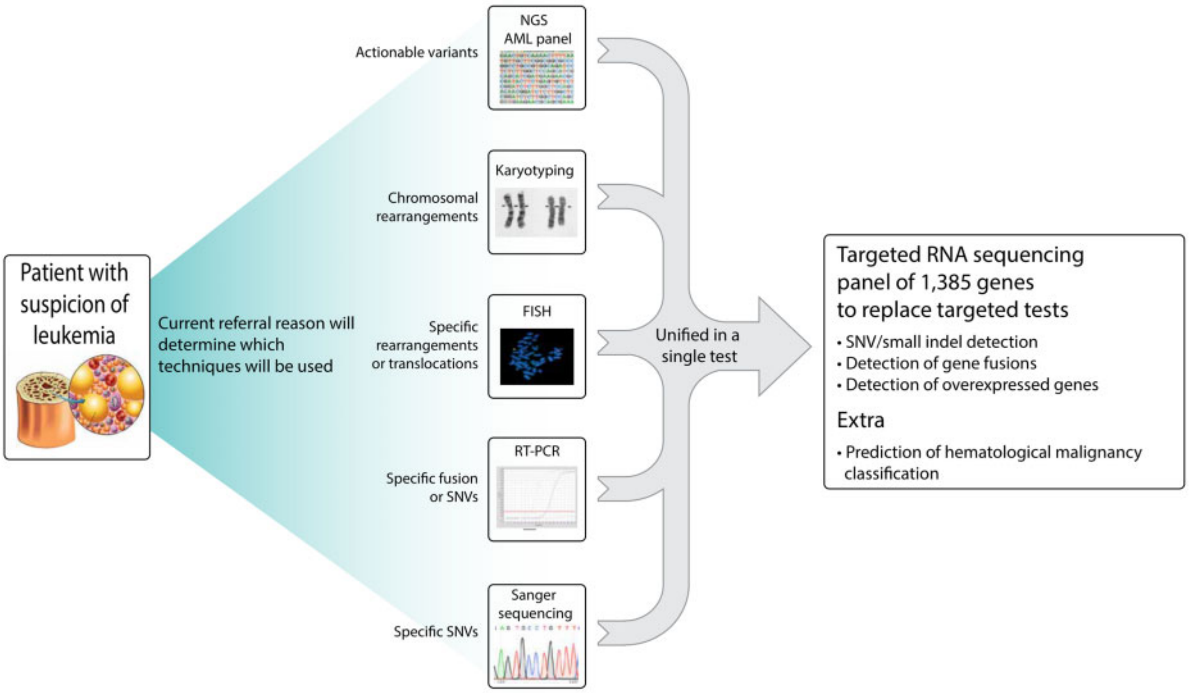
Patient with suspicion of leukemia
Current referral reason will determine which techniques will be used
Actionable variants
NGS AML panel
Chromosomal rearrangements
Karyotyping
Specific rearrangements or translocations
FISH
Specific fusion or SNVs
RT-PCR
Specific SNVs
Sanger sequencing
Unified in a single test
Targeted RNA sequencing panel of 1,385 genes to replace targeted tests
• SNV/small indel detection
• Detection of gene fusions
• Detection of overexpressed genes
Extra
• Prediction of hematological malignancy classification

Fig. 1. Schematic overview of which techniques can be replaced with a targeted RNA-sequencing approach. Based on the referral type of hematological malignancy, one technique (or a combination) is currently used to detect genetic abnormalities relevant for prognosis and treatment. This overview shows which techniques can be replaced by a single targeted RNA-sequencing test. Note that other techniques to identify and classify hematological malignancies are not shown.

the variance per component, we developed a decision tree that placed samples into 3 HM categories: AML, ALL, and an "unclassified" category containing myelodysplastic syndromes (MDS), myeloproliferative neoplasms (MPN), chronic myeloid leukemias (CML), and lymphoma. As validation set, we used 60 samples. We then measured the prediction performance using an additional application set of 24 samples. Expression of *EVI1* was measured for samples in the AML category. For the unclassified category, including lymphoma samples, we quantified *CCND1* and *BCL2* expression.

SAMPLE PREPARATION, SEQUENCING, AND DATA ANALYSIS

RNA was converted into cDNA, and libraries were generated using the Illumina TruSight RNA Pan-Cancer panel (RS-303-1002, Illumina), following the manufacturer's protocol. The quality and quantity of the libraries was measured using TapeStation D1000 (Agilent). After the measurement libraries were equimolarly pooled, there were 24 samples per pool for a NextSeq run and 8 samples for a MiSeq run. Pools were paired-end sequenced 2 × 75 bp (6 bp index) cycles on a NextSeq550 or MiSeq instrument using the NextSeq Mid output kit V2 (Illumina, FC-404-2001) or the MiSeq V3 kit (Illumina, MS-102-3001), according to manufacturer's protocol, generating between 5 and 8 million reads. NextSeq data were demultiplexed using our in-house tool (https://github.com/molgenis/NGS_Demultiplexing). MiSeq data were automatically demultiplexed on the sequencer using MiSeq reporter.

Demultiplexed fastq files were uploaded to the Basespace analysis environment (Illumina) and analyzed with the RNA-seq alignment app (v.1.0.0) using default settings for fusion calling with RefSeq hg19 gene annotation. The fusion calls were detected with MANTA (https://github.com/Illumina/manta), a tool that uses paired-end sequencing reads to call structural variants and medium-sized indels. In addition, SNPs and small indels were annotated with the Isaac variant caller and collected in variant call format (VCF) files within the Basespace RNA-seq alignment app.

Expression data was generated using our in-house RNA-seq pipeline (https://github.com/molgenis/NGS_RNA). Alignment was performed using STAR

Table 1. Variant detection from targeted RNA-sequencing testing[a].

Variant detection	Genes investigated	Leukemia classification	Validation set (no. of samples)	Application set (no. of samples)
SNV/indel	Myeloid panel 54 genes	AML	63	
		MDS		13
	Lymphoid panel 72 genes	ALL		14
Total number			63	27
Fusion	Pan-cancer panel 1385 genes	ALL/AML/MDS/ MPN/CML/ lymphoma	40	96
Total number			40	96
Overexpression	*EVI1*	AML	24	6
		ALL	6	5
	BCL2, CCND1	Unclassified category (including MDS/ MPN/ CML/lymphoma)	30	13
Total number			60	24

[a]The sample cohort was split into a validation and an application set. For SNV/indel detection, AML samples ($N = 63$) were chosen for the validation set when results from current techniques were available. MDS and ALL samples out of the cohort were chosen as application set. For fusion detection, the first 40 samples out of the total cohort were used as validation set and the other 96 samples as application set. For overexpression, 84 samples with unambiguous leukemia classification were used. Of these, the first 60 samples out of the cohort were used as validation set and the other 24 samples used as the application set.

alignment software (https://www.ncbi.nlm.nih.gov/pubmed/23104886) with the same settings as the RNA-seq alignment app. HTSeq (v.0.9.1, https://github.com/simon-anders/htseq) read counts were filtered on all 1385 genes in the TruSeq RNA panel and normalized to remove technical bias. These counts were then quantile-normalized to the median distribution and $\log_2$ transformed to reduce noise. Sample means were centered to zero.

SNV/INDEL CALLING

The VCF files from the Isaac variant caller were further analyzed using Alissa Interpret (Agilent). Two virtual gene panels, myeloid and lymphoid, were assembled for filtering. The genes in both panels were selected according to the TrueSight Myeloid sequencing panel (Illumina) for AML-MDS and ALL, based on literature (8–10) (see Supplemental Table 1 for a list of analyzed genes).

Variants present in ClinVar (11), HGMD (12) or COSMIC (13) were considered relevant. All other variants were filtered on read depth (>15), population frequency (≥2%), allele frequency, and ≥400 allele count from the Exome Aggregation Consortium (14) or the genome aggregation database (15) and Genome of the Netherlands (16). After filtering, remaining variants were subjected to further interpretation. Variants classified as variants of unknown significance (VUS), likely pathogenic and pathogenic were reported. If a variant was present in the COSMIC database, its pathogenicity was checked in the HM cohort in COSMIC and classified accordingly. If a variant was not present in COSMIC, it was evaluated for its potential pathogenicity using in silico prediction tools and data available in the Alissa clinical informatics database from Agilent and Alamut software (v.2.8–2.11; Interactive Biosoftware). If at least 3 out of the 5 tools in Alamut predicted a damaging or pathogenic effect, including a splice site effect, the variant was classified as VUS, otherwise it was classified as (likely) benign.

FUSION GENE CALLING

Each potential fusion gene was considered when it met a quality score of 0.6 from MANTA, we detected at least one paired and one split, and a minimal fusion read number of 10 unique fusion reads or wild type reads were present for both genes of the fusion. Fusion reads were omitted when the complete fusion read could be mapped elsewhere in the human genome due to pseudogenes or paralogs.

OVEREXPRESSION DETECTION

PCA was performed using the prcomp function in R (v.3.5.2) (17). First, to exclude technical variation due to RNA isolation batch, sample preparation, and sequence run, the sample variation of these components was measured. Second, samples were labeled based on the differential diagnosis classification of leukemia and split into a validation and an application set. Classification of HM samples was performed with rpart (v.4.1–13) (18) using the first 5 principal components of the validation set, which yielded 3 classification categories: ALL, AML, and unclassified. The unclassified category includes MDS, MPN, CML, and lymphoma samples. We called this process RANKING, and benchmarked it using the application set, see https://github.com/kdelange/RANKING for the R scripts used.

Expression levels were measured for *EVI1* in AML and for *CCND1* and *BCL2* in the unclassified category. The relative expression is shown in boxplots made using the ggplot2 packages (19) for genes of interest per HM type. Overexpression was defined as expression significantly outside the normal expression distribution ($P < 0.05$) within the same category.

CONFIRMATION OF ADDITIONAL VARIANTS DETECTED WITH TARGETED RNA-SEQUENCING

Additional SNVs and/or indels were confirmed with Sanger sequencing for variants with variant allele frequency (VAF) $> 10\%$ or with MLPA or ddPCR for variants with VAF $< 10\%$. Additional fusion genes were confirmed with PCR on cDNA and, if possible, FISH. See online Supplemental Data for the materials, methods, and primers and probes (Supplemental Table 2).

ROUTINE (CYTO)GENETIC METHODS

Karyotyping and additional FISH were performed according to Dutch national guidelines (20). SNP-array profiling was performed on cases referred for MDS using the Illumina Infinium® Global Screening Array platform and analyzed with Nexus/NxClinical software (BioDiscovery). For cases referred for AML, RT-PCR was performed for *KIT* and *NPN1*, and Sanger sequencing for *CEBPA*. NGS was performed using the TrueSight Myeloid sequencing panel (Illumina). See Supplemental Materials and Methods for details.

Results

SNV/INDEL DETECTION

In total, 113 variants were detected in 90 patients. Of these, confirmation was performed with at least one independent method for 83 variants: 60 variants from the validation set and 23 from the application set. The remaining 30 variants were classified as VUS, and no confirmation was performed (Table 2, Supplemental Table 3).

In total, 82 of the 83 variants identified using targeted RNA-sequencing were in agreement with the results from current methods, resulting in 98.8% sensitivity. Targeted RNA-sequencing missed one variant (c.2447A>T; D816V13) in the *KIT* gene detected by RT-PCR at 10^{-4}. In 32 samples, no relevant variants were detected [AML amplicon panel ($N = 7$) and RT-PCR ($N = 25$)]. Curiously, the AML amplicon panel

Table 2. Comparison of clinically relevant variants detected with current methods and with targeted RNA sequencing[a].

Current methods compared with RNA-seq	No. of samples	No. of analyzed genes	No. of variants detected	TP	FP	FN	Sensitivity (%)
SNV/ indel	90	54 [myeloid panel] 72 [lymphoid panel]	PC: 83 (+ 30 VUS)	82	0	1	98.8
			CM: 83	83	0	0	100
Gene fusions	136	1385	PC: 33	33	0	0	100
			CM: 33	27	0	6	81.8
Gene over expression	84	3	PC: 11	10	0	1	90.9
			CM: 11	9	0	2	81.8

TP = true positive, FP = false positive, FN = false negative, PC = targeted pan-cancer RNA-sequencing, CM = current methods.

[a] The SNV/indel validation set ($N = 63$) consists of 28 samples previously sequenced with an amplicon myeloid panel of 54 genes and 35 samples where either RT-PCR or Sanger sequencing was performed for the *CEBPA*, *KIT*, and *NPM1* genes. In total, 59 clinical relevant variants were detected. In the application set, 23 clinical relevant variants were detected and confirmed with a second technique. Compared to current techniques, our method missed just one variant: **KIT*: D816V13 detected by RT-PCR at 10^{-4}.

For fusion genes, our method detected 33 variants, of which 27 had been identified by current methods. The 6 additional gene fusions detected by our method, *AUTS2/PAX5*, *EBF2/PDGFRB*, *NUP98/PSIP1*, *NUP214/ABL*, *TGF/GPR128*, and *NUP98/SET*, were then confirmed using a second technique.

For overexpression of the genes *EVI1*, *CCND1*, and *BCL2*, 10 samples showed overexpression: 7 cases of *EVI1* in and 3 cases of *CCND1*. Compared to current techniques, our method missed only one overexpressed gene, *BCL2***, for which the current FISH test showed *IGH-BCL2* t(14; 18)(q32; q21) in 2% of the cells. Of the 10 samples with overexpression, 2 samples showed overexpression for *CCND1* that was confirmed as a t(11; 14) with interphase FISH.

detected 4 variants that were not identified using targeted RNA-sequencing. Two of these had a low VAF after AML amplicon testing, 0.06 (*RUNX1*) and 0.05 (*BCOR*), and could not be confirmed by additional validation experiments with specific MLPA primers, which means they are most likely artifacts. The other 2 variants had VAFs of 0.11 (*NOTCH*) and 0.16 (*DNMT3A*) and were not confirmed by Sanger sequencing, meaning they are most likely false positives of the amplicon sequencing and would not even be reported given the recent adoption of more stringent reporting criteria.

DETECTION OF FUSION GENES

Of the 136 samples, 2 samples failed because not enough input material was available. Targeted RNA-sequencing detected 27 fusion genes in 27 samples, in agreement with the results of current methods. In the other 107 samples, both targeted RNA-sequencing and current diagnostics detected no fusion genes, reaching 100% sensitivity.

Out of the 27 fusion genes detected, 11 different types of fusion genes (consisting of different genes or having different breakpoints) were found (Table 3, Supplemental Table 4). Targeted RNA-sequencing also detected 6 gene fusions in 5 different samples that had not been found with current methods.

USING RANKING TO PREDICT LEUKEMIA CLASSIFICATION BASED ON EXPRESSION PROFILING

We selected the first 5 principal components from the expression-based PCA, which explain 45% of the variance, and observed that samples with the same HM classification clustered together (Fig. 2, Supplemental Figs. 2 and 3, Supplemental Table 4). RANKING then used the 5 principal components to automatically HM samples. This analysis classified one CML sample as AML, but our re-examination of this sample revealed it to be an AML sample. We initially used 6 classes in RANKING (AML, ALL, MPN, MDS, CML, and lymphoma), and found that prediction was correct in approximately 60% of all samples (Supplemental Table 4). While prediction was 100% correct for AML and ALL, performance was much lower for the other classes. We therefore chose to use only 3 classifications in RANKING: AML, ALL, and an ‘unclassified’ group. These 3 categories were then used to predict HM classification in the independent application set, and prediction was 100% correct for these test samples [AML ($N = 6$), ALL ($N = 5$), and unclassified ($N = 13$)] (Fig. 3).

DETECTION OF OVEREXPRESSION

Overexpression of *EVI1* was measured in AML samples. According to the type of HM, 30 AML samples were present in our cohort and 7 samples showed overexpression (Supplemental Fig. 4, A), in agreement with the results of current diagnostics using RT-PCR or FISH. We detected significant *EVI1* overexpression in *EVI1*-positive samples compared to the other AML samples (P-values $\leq 1.75 \times 10^{-5}$).

Overexpression was measured for *CCND1* and *BCL2* in the unclassified sample category, which includes lymphoma samples. Three samples showed *CCND1* overexpression (Supplemental Fig. 4, B) as compared to unclassified samples without *CCND1* overexpression. All 3 positive samples had significant P-values (3.11×10^{-7}, 2.18×10^{-7}, and 1.03×10^{-28}). For one sample this was in agreement with previous FISH results. For the others, *CCND1* overexpression was confirmed with FISH, which showed a translocation (11; 14)(q13; q32)

Table 3. Overview of all detected gene fusions.

Fusion gene	Breakpoints	Translocation	Primary detected with
BCR/ABL	chr22:23,632,600, chr9:133,729,451	t(9; 22)(q34.12; q11.23)	K, F
PML/RARA	chr15:74,315,747, chr17:38,504,566	t(15; 17)(q24.1; q21.2)	K, F, RT-PCR
CBFB/MYH11	chr16:67,116,209, chr16:15,814,906	inv(16)(p13.11q22.1)	K; F
KMT2A/MLLT1	chr11:118,355,688, chr19:6,270,770	t(11; 19)(q23.3; p13.3)	F
NUP214/ ABL	chr9:134,106,154, chr9:133,729,449	t(9; 9)(q34.13; q34.12)	PC panel
SET/ NUP98	chr9:131,453,448, chr11:3,781,769	t(9; 11)(q34.11; p15.4)	PC panel
PDGFRB/ EBF1	chr5:149,506,177, chr5:158,134,987	t(5; 5)(q32; q33.3)	PC panel
TGF/GPR128 (CCF9)	chr3: 100,438,901, chr3:100,348,441	t(3; 3)(q12.2.1; q12.2)	PC panel
PAX5/ AUTS2	chr9:36,966,545, chr7:70,163,552	t(7; 9)(q11.2; p13.2)	PC panel
PSIP1/ NUP98	chr9:15,479,684, chr11:3,784,131	t(9; 11)(p22.3; p15.4)	PC panel

K = karyotyping, F = FISH, PC panel = targeted pan-cancer RNA-sequencing.

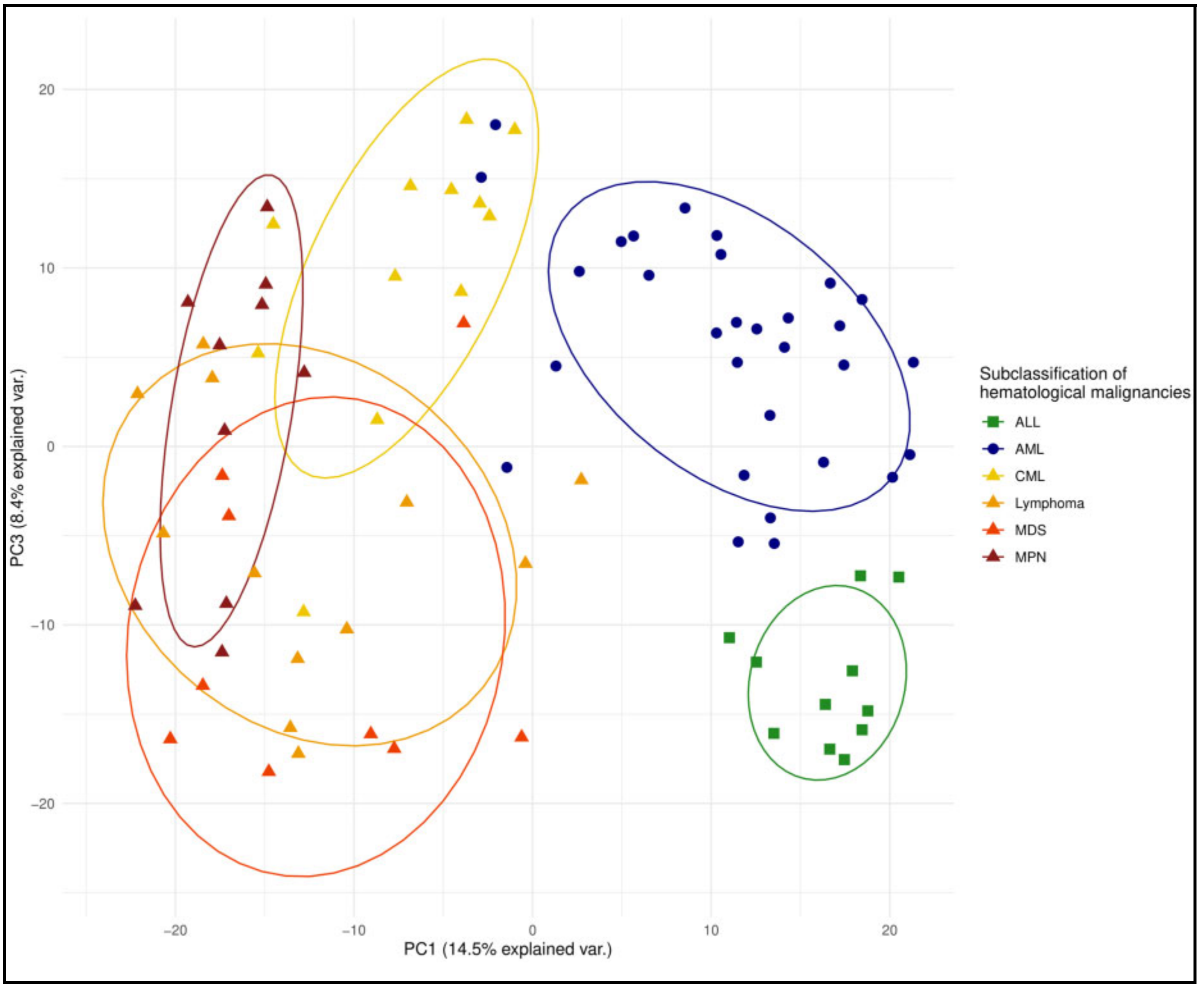

Fig. 2. Principal component analysis for hematological malignancy classification. Figure shows principal component 1 (PC1), which explains 14.5% of the variance, versus principal component 3 (PC3). For our analyses we started with 6 different types of hematological malignancies.

that had not been detected by the current diagnostic approach. In one case, our approach did not detect overexpression of *BCL2* (Supplemental Fig. 3, C). Karyotyping and additional FISH showed a t(14; 18)(q32; q21)(*IGH-BCL2*) in less than 2% of the cells.

Discussion

We have shown that targeted RNA-sequencing is a feasible method for genetic testing of different types of HM. The combination of targeted RNA-sequencing and the workflow we developed can replace and improve upon the current routine molecular (cytogenetic) tests used to detect relevant genetic abnormalities. Our approach can assist in the assignment of the correct diagnosis of HM classification.

Applying a decision tree to the principal components of targeted expression to predict HM classification yielded promising results. Our prediction tool, RANKING, showed an accuracy of 100% for all the AML and ALL samples in the application set. We observed that the classification of AML and ALL is the most reliable, and that classification of other HM types in our dataset is more challenging. Therefore, we defined an 'unclassified' category for the other types of HM included in this study. However, this does not mean that they have similar expression patterns. Larger datasets, including healthy bone marrow samples as control, will help to discriminate between MPN, MDS, and lymphomas. Thus, even though RANKING's predictions are not perfect, they can guide classification. In our validation set, we found 3 samples that did not fit

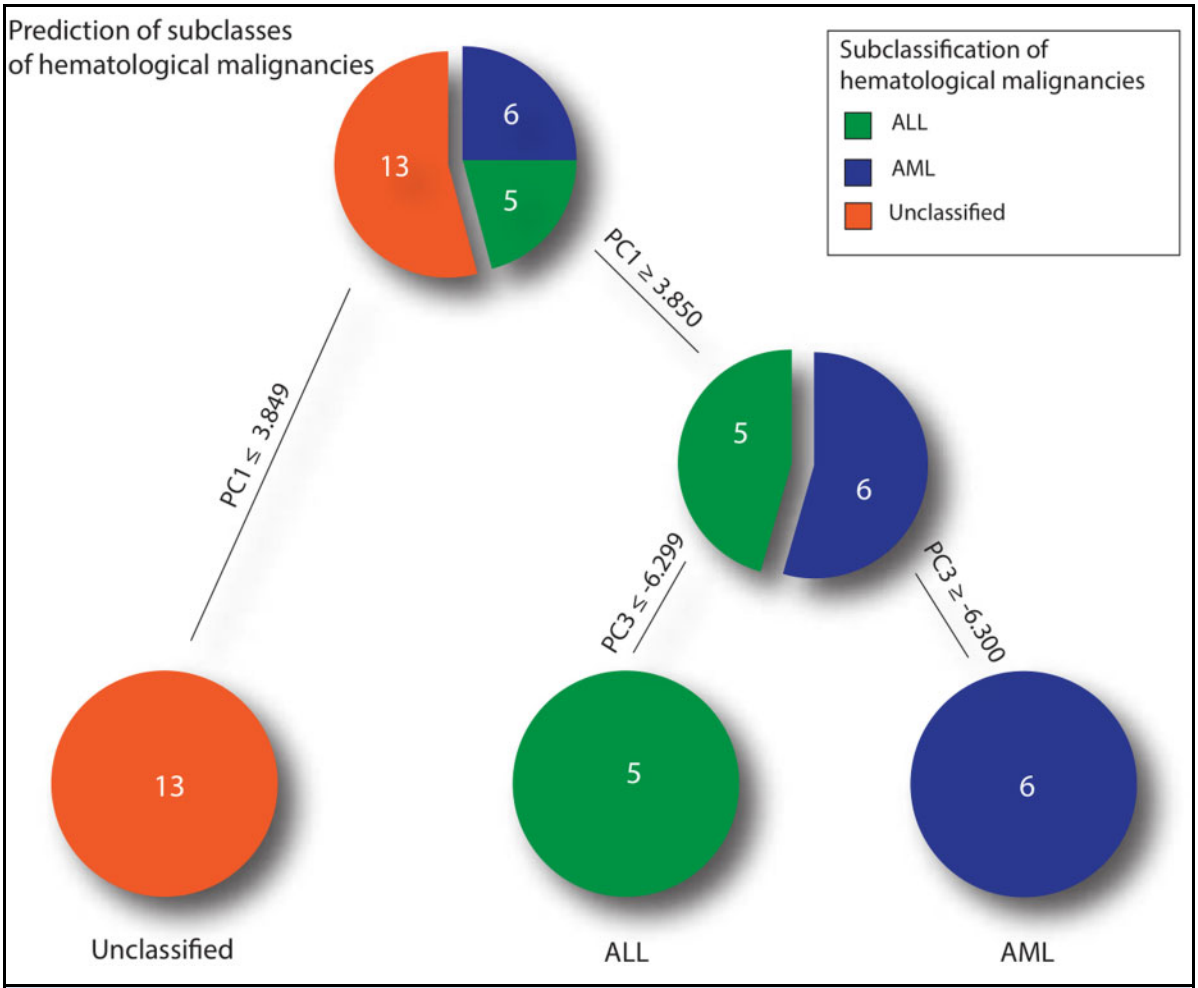

Fig. 3. Result of the decision tree using the application set. Three categories were defined: AML, ALL, and unclassified. The unclassified category contains MDS, MPN, CML, and lymphoma samples.

the profile of the referred type of HM. Careful reanalysis of these cases revealed that our prediction was correct. Moreover, previous methods have demonstrated the potential of using gene-expression information to classify HMs (4). In future work, we also aim to identify aberrant splicing, an approach has been shown to have added value for Mendelian diseases (21). We thus expect that, with larger transcriptome sequencing datasets, it will also become possible to more reliably predict the other HM types.

We were able to detect SNV/indels with 98.8% sensitivity compared to current methods, and our method only missed one variant detected at 10^{-4} by RT-PCR (c.2447A>T; D816V13, *KIT*). In general, RNA-based methods have the advantage that they are enriched for highly expressed transcripts and their variants, which are more likely to be pathogenic. A disadvantage of an RNA-sequencing approach for SNV/indel detection could be the discovery of variants that destabilize transcripts (nonsense-mediated decay) or variants in tumor-suppressor genes (22) or genes having mono-allelic expression. However, we did not observe any of these in the present study.

Our workflow can detect translocations resulting in gene fusions with any of the 1385 genes included in the panel, with any breakpoint. This allows for the detection of relevant fusions irrespective of the referral type of HM. Translocations, often resulting in gene fusions, are typically identified using FISH or RT-PCR where one or a few potential translocations are tested based on the referral type of leukemia. Other targeted RNA assays were developed for fusion gene detection (23–25). While these assays have a high sensitivity and specificity, they are designed for fusion gene detection only. In our method, all fusion genes were detected with 100% accuracy in a single experiment. We even found 6 fusion genes using RNA-sequencing that had not been found by other methods, and 4 of these genes were relevant for diagnosis and risk prediction.

Translocations detectable with FISH or karyotyping do not necessarily result in fusion genes. One example is the translocation or inversion of the *EVI1* gene caused by inv(3)(q21q26) or t(3; 3)(q21; q26), which results in a very poor prognosis in AML (26) and leads to overexpression of EVI1. In our cohort, our method detected EVI1 overexpression confirming karyotyping or FISH results in all 7 cases. In addition, in translocations including the *IGH* gene, no fusion gene is detectable at RNA-level. In our cohort, we detected overexpression of *CCND1* in 3 lymphoma samples, 2 of them found first by RNA-sequencing. More samples are needed to determine the minimum number of aberrant cells for different breakpoints required to measure overexpression of relevant genes or to measure loss of gene expression due to translocations. However, our results are promising, especially with respect to detection of *IGH* rearrangements given that the diversity of the rearrangements make current methods laborious.

Currently, we are unable to detect CNVs with our targeted RNA-sequencing method. While it is possible to detect CNVs using transcriptomics, this requires large training datasets (27). For now, we recommend users combine our test with a simple genotyping array chip. This will allow genome-wide detection of copy numbers and hypo- and hyperdiploidy. A combination of targeted RNA-sequencing and genotype array will allow for the identification of all genetic aberrations relevant to HMs.

Another shortcoming of our approach is the detection of tandem duplications. Reliable detection of the common internal tandem duplication in exon 14-15 of *FLT3,* for instance, is of special importance in AML subclassification since presence of the *FLT3-ITD* is an indication for the use of tyrosine kinase inhibitors (28). In our approach, however, we were not able to detect this duplication. This is a known problem of the existing software tools (29) due to alignment problems. However, a recently developed tool called ReSCU uses soft clipped reads to detect duplications in *FLT3* and *KMT2A* (6), which might overcome this problem.

In conclusion, we performed targeted RNA-sequencing and detected relevant fusion genes, SNVs, and overexpressed genes with high accuracy compared to current genetic techniques. We then predicted the HM classification based on expression profiling. Our streamlined workflow can replace and improve upon current routine molecular (cytogenetic) tests to detect relevant genetic abnormalities but needs to be tested in a large prognostic cohort to establish its diagnostic value.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Human Genes: *EVI1*, ecotropic viral integration site 1 *CCND1*, cyclin D1 *BCL2*, BCL2 apoptosis regulator

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. K. de Lange, statistical analysis, administrative support; A. Bosga-Bouwer, provision of study material or patients; A. Mulder, provision of study material or patients.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: ZONMW, grant no. 40-41200-98-9159.
Expert Testimony: None declared.
Patents: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

Acknowledgments: We thank Kate McIntyre for editorial advice, Gerben de Vries for installing the RNA pipeline, Pei Ming for developing and use of ddPCR primers and René Mulder, Arjan Simpelaar, and Bettie Postema-Veenema for performing routine NGS, RT-PCR (*KIT* and *NPM1*) and Sanger sequencing (*CEBPA*).

References

1. Duncavage EJ, Abel HJ, Szankasi P, Kelley TW, Pfeifer JD. Targeted next generation sequencing of clinically significant gene mutations and translocations in leukemia. Mod Pathol 2012;25:795–804.
2. Heim S, F M. Cancer cytogenetics. 4th Ed. Wiley Blackwell; 2015. United Kingdom
3. Dougherty MJ, Wilmoth DM, Tooke LS, Shaikh TH, Gai X, Hakonarson H, et al. Implementation of high resolution single nucleotide polymorphism array analysis as a clinical test for patients with hematologic malignancies. Cancer Genet 2011;204:26–38.
4. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
5. Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer 2015;15:371–81.
6. Arindrarto W, Borràs DM, de Groen RAL, van den Berg RR, Locher IJ, van Diessen SAME, et al. Comprehensive

diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing. Leukemia 2020; doi:. 10.1038/s41375-020-0762-8.

7. Ringnér M. What is principal component analysis? Nat Biotechnol 2008;26:303–4.
8. Iacobucci I, Mullighan CG. Genetic basis of acute lymphoblastic leukemia. J Clin Oncol 2017;35:975–83.
9. Sujobert P, Le Bris Y, de Leval L, Gros A, Merlio JP, Pastoret C, et al. The need for a consensus next-generation sequencing panel for mature lymphoid malignancies. Hemasphere 2019;3:e169.
10. Reshmi SC, Harvey RC, Roberts KG, Stonerock E, Smith A, Jenkins H, et al. Targetable kinase gene fusions in high-risk B-ALL: a study from the Children's Oncology Group. Blood 2017;129:3352–61.
11. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46:D1062–7.
12. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 2017; 136:665–77.
13. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004;91:355–8.
14. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285–91.
15. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv 2019.
16. Francioli LC, Menelaou A, Pulit SL, Van Dijk F, Palamara PF, Elbers CC, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 2014;46:818–25.
17. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing 2018.
18. Therneau T, Atkinson B, Ripley B, Ripley MB. rpart: Recursive partitioning and regression trees 2015. Available at https://rdrr.io/cran/rpart/ (Accessed August 2020).
19. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag; 2016.
20. Snijder S, Berg-de Ruiter Evd, Beverloo B, Arjan BJJ, van der Kevie-Kersemaekers A-M, Knijnenburg J, et al. VKGL kwaliteitscommissie_veldnorm titel: Richtlijnen verworven cytogenetica Doc. Code. Genoomdiagnostiek-Cytogenetica: VKGL_V07 Subspecialisme; 2017. https://www.vkgl.nl/nl/diagnostiek/formulieren-documenten-kwaliteit/category/7-veldnormen (Accessed August 2020).
21. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med 2017; 9(386):eaal5209..
22. White BS, DiPersio JF. Genomic tools in acute myeloid leukemia: from the bench to the bedside. Cancer 2014; 8:1134–44.
23. Zheng Z, Liebers M, Zhelyazkova B, Cao Y, Panditi D, Lynch KD, et al. Anchored multiplex PCR for targeted next-generation sequencing. Nat Med 2014;20: 1479–84.
24. Dillon LW, Hayati S, Roloff GW, Tunc I, Pirooznia M, Mitrofanova A, et al. Targeted RNA-sequencing for the quantification of measurable residual disease in acute myeloid leukemia. Haematologica 2019;104:297–304.
25. Patkar N, Bhanshe P, Rajpal S, Joshi S, Chaudhary S, Chatterjee G, et al. NARASIMHA: novel assay based on targeted RNA sequencing to identify chimeric gene fusions in hematological malignancies. Blood Cancer J 2020;10:1–4.
26. Hinai AA, Valk PJM. Review: aberrant EVI1 expression in acute myeloid leukaemia. Br J Haematol 2016;172: 870–8.
27. Fehrmann RSN, Karjalainen JM, Krajewska M, Westra HJ, Maloney D, Simeonov A, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet 2015;47:115–25.
28. Stone RM, Mandrekar SJ, Sanford BL, Laumann K, Geyer S, Bloomfield CD, et al. Midostaurin in FLT3-mutated acute myeloid leukemia: the authors reply. N Engl J Med 2017;377:454–64.
29. Prieto-Conde MI, Corchete LA, García-Álvarez M, Jiménez C, Medina A, Balanzategui A, et al. A new next-generation sequencing strategy for the simultaneous analysis of mutations and chromosomal rearrangements at DNA level in acute myeloid leukemia patients. J Mol Diagnostics 2020;22:60–71.