


## RESEARCH ARTICLE

# The transcriptome-wide landscape of molecular subtype-specific mRNA expression profiles in acute myeloid leukemia

Tian Mou<sup>1,2</sup>  | Yudi Pawitan<sup>1</sup> | Matthias Stahl<sup>3</sup> | Mattias Vesterlund<sup>3</sup> | Wenjiang Deng<sup>1</sup> | Rozbeh Jafari<sup>3</sup> | Anna Bohlin<sup>4</sup> | Albin Österroos<sup>5</sup> | Ioannis Siavelis<sup>3</sup> | Helena Bäckvall<sup>3</sup> | Tom Erkers<sup>3</sup> | Santeri Kiviluoto<sup>3</sup> | Brinton Seashore-Ludlow<sup>3</sup> | Päivi Östling<sup>3,6</sup> | Lukas M. Orre<sup>3</sup> | Olli Kallioniemi<sup>3,6</sup> | Sören Lehmann<sup>4,5</sup> | Janne Lehtiö<sup>3</sup> | Trung Nghia Vu<sup>1</sup>

<sup>1</sup>Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

<sup>2</sup>School of Biomedical Engineering, Shenzhen University, Shenzhen, China

<sup>3</sup>Department of Oncology Pathology, Karolinska Institutet, Science for Life Laboratory, Stockholm, Sweden

<sup>4</sup>Department of Medicine Huddinge, Karolinska Institutet, Unit for Hematology, Karolinska University Hospital Huddinge, Stockholm, Sweden

<sup>5</sup>Department of Medical Sciences, Section of Hematology, Uppsala University Hospital, Uppsala, Sweden

<sup>6</sup>Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

## Correspondence

Trung Nghia Vu, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solnavägen 1, 171 77 Solna, Sweden.

Email: trungnghia.vu@ki.se

## Funding information

Cancerfonden; Barncancerfonden; Knut och Alice Wallenbergs Stiftelse; Stiftelsen för Strategisk Forskning; Vetenskapsrådet

## Abstract

Molecular classification of acute myeloid leukemia (AML) aids prognostic stratification and clinical management. Our aim in this study is to identify transcriptome-wide mRNAs that are specific to each of the molecular subtypes of AML. We analyzed RNA-sequencing data of 955 AML samples from three cohorts, including the BeatAML project, the Cancer Genome Atlas, and a cohort of Swedish patients to provide a comprehensive transcriptome-wide view of subtype-specific mRNA expression. We identified 729 subtype-specific mRNAs, discovered in the BeatAML project and validated in the other two cohorts. Using unique proteomics data, we also validated the presence of subtype-specific mRNAs at the protein level, yielding a rich collection of potential protein-based biomarkers for the AML community. To enable the exploration of subtype-specific mRNA expression by the broader scientific community, we provide an interactive resource to the public.

## 1 | INTRODUCTION

Acute myeloid leukemia (AML) is a heterogeneous disease due to the diversity of genetic alterations.<sup>1,2</sup> These alterations determine AML pathophysiology, progression and response to therapy, and thus disease heterogeneity complicates clinical management.<sup>3,4</sup> Many studies have aimed to stratify AML patients into molecular subtypes, each

with distinct clinical significance in terms of prognosis or therapy response.<sup>5-7</sup> For example, favorable outcomes are reported for patients with fusion subtypes, such as *PML-RARA* and *CBFB-MYH11*, as well as the non-fusion subtypes, such as *CEBPA*<sup>Biallelic</sup> subtype.<sup>6,8</sup> On the other hand, overexpression of *EVI1* is a poor prognostic factor in *MLL*-rearranged AML.<sup>9</sup> A fully genomic classification of AML that considers patterns of co-mutations has been established recently

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2021 The Authors. *American Journal of Hematology* published by Wiley Periodicals LLC.

based on a large patient cohort.<sup>6</sup> At least 11 molecular subtypes can be recognized based on cytogenetic and targeted sequencing analyses. However, large-scale information on how these genetic differences manifest themselves at the transcriptome and proteome level is not yet established. Therefore, our aim is to identify transcriptome-wide mRNAs that are specific to each of the molecular subtypes, and to validate their significance at the protein level. Subtype-specific mRNAs or the corresponding protein can provide further biological understanding and be utilized as potential biomarkers for diagnosis, classification, therapy response, as well as disease monitoring.

Here, we use a novel method that leverages available large-scale RNA-seq datasets to identify molecular subtype-specific mRNA expression patterns. The specificity of the genes is established using two statistics, one for testing the genes that are over-expressed in a single subtype compared to all the other subtypes and the other for testing whether the remaining subtypes are statistically different.

We analyzed next-generation RNA-sequencing (RNA-seq) data of 955 AML samples from three cohorts, including the BeatAML project (N = 461),<sup>10</sup> The Cancer Genomic Atlas (N = 179)<sup>2</sup> and a cohort of Swedish patients (N = 315),<sup>11</sup> called the Clinseq cohort from here on. To provide these results and findings to the broader scientific community, we have created an interactive resource containing the complete results of subtype-specific analysis at <https://nghiaivr.shinyapps.io/AMLSubtypeSpecificDiscovery/>.

## 2 | MATERIALS AND METHODS

### 2.1 | Study cohorts

The statistical identification of subtype-specific genes is optimized using several tuning parameters, including the choice of contrast statistics and their corresponding thresholds. The BeatAML cohort which contains 461 AML patients is used as a discovery cohort to determine subtype-specific genes. The TCGA cohort (179 samples), and the Clinseq cohort (315 samples) are used for validation. The details of each cohort are described below. The FASTQ files were obtained from the three cohorts and RNA-seq reads were aligned to the human genome hg19. Also, XAEM<sup>12</sup> was used to obtain the gene expression (transcript per million - TPM) from the RNA-seq data. The calculation process of XAEM followed the instructions provided at <http://fafner.meb.ki.se/biostatwiki/xaem/>.

#### 2.1.1 | BeatAML data

The BeatAML project generated functional genomic data of primary bone marrow biopsies from patients with AML. The dataset includes genomic and transcriptomic analyses, clinical annotations and drug responses. Patients in this cohort have received standard intensive chemotherapy analogous to TCGA and Clinseq cohorts. Samples were first processed with the Agilent SureSelect Strand-Specific RNA Library Preparation Kit on the Bravo robot (Agilent). Sequencing was

performed on the Illumina HiSeq 2500 platform using a 100-cycle paired-end protocol. More details of the data can be found in the original paper.<sup>10</sup> In total 23 360 genes across 461 samples with complete clinical information and drug response are used in this study.

#### 2.1.2 | TCGA data

The systematic study of the Cancer Genome Atlas (TCGA) AML samples has provided a genomic landscape of AML and generated a catalogue of leukemia-related genes.<sup>2</sup> There is, therefore, a possibility to also make use of this sequencing data for a more refined understanding of subtype-specific patterns of mRNA expression. RNA-sequencing was performed using Illumina HiSeq 2000 PE 75 base sequencing protocol. Patients of the TCGA-AML study received intensive induction treatment (chemotherapy). A total of 22 374 genes from the 179 samples were used in this study as a validation set.

#### 2.1.3 | Clinseq data

The Clinseq cohort includes 315 patients diagnosed with AML in Sweden between February 1997 and August 2014.<sup>11</sup> Clinical information was retrieved from patient records and the Swedish Adult Acute Leukemia Registry. All patients underwent intensive induction therapy (including anthracyclines and cytosine arabinoside) as first-line treatment. Bone marrow or peripheral blood samples were obtained at the time of diagnosis, separated for mononuclear cells and stored at  $-180^{\circ}\text{C}$  until use. Transcriptomic RNA was sequenced using the Illumina HiSeq-2500 platform. As a validation set, gene expressions of 23 572 genes across 315 samples are considered in the analysis.

A representative subset of 118 patient samples with sufficient biological material was selected from the original cohort for mass spectrometry-based proteomics analysis. The corresponding protein levels of 12 142 genes were quantified in at least one sample. Mass spectrometry based proteomics was carried out as previously described.<sup>13-15</sup> Briefly, viably frozen patient samples were washed, and the cells were lysed by 4% SDS lysis buffer and prepared for mass spectrometry analysis using a modified version of the SP3 protein clean up and digestion protocol.<sup>16</sup> Peptides were labeled with TMT10-plex reagent according to the manufacturer's protocol (Thermo Fisher Scientific) and separated by immobilized pH gradient–isoelectric focusing (IPG-IEF) on 3–10 strips as described previously.<sup>13</sup> Extracted peptide fractions from the IPG-IEF were separated using an online 3000 RSLCnano system coupled to a Thermo Fisher Scientific Q Exactive-HF. MSGF+ and Percolator in the Nextflow platform were used to match MS spectra to the Ensembl92 human protein database.<sup>17,18</sup>

### 2.2 | Molecular-subtype classification of AML

A genomic classification of AML, proposed by Papaemmanuil et al.,<sup>6</sup> is used in this study for the identification of subtype-specific genes.

Papaemmanuil's classification compartmentalized AML into 11 subtypes, each with distinct diagnostic features and considerable relevance to clinical outcomes. Although the 11 subtypes well reflect the genetic characteristics of AML, the association between molecular mutations and the gene expression level are not well described.

We applied this genomic classification to the BeatAML, TCGA and Clinseq cohorts and found similar frequencies of membership in each subtype, shown in Table S1. For simplicity, in this context, we named each subtype as follows:

- *NPM1*—AML with *NPM1* mutation;
- *TP53*-mutant—AML with *TP53* mutation, chromosomal aneuploidy or both;
- Splice—AML with mutated chromatin, RNA-splicing genes, or both;
- *CBFB-MYH11*—AML with *inv(16)(p13.1;q22)* or *t(16;16)(p13.1;q22)*;
- *PML-RARA*—AML with *t(15;17)(q22;q12)*;
- *MLL*—AML with *MLL* fusion genes, *t(x;11)(x;q23)* multiple fusion partners for *MLL*;
- *CEBPA*<sup>Biallelic</sup>—AML with biallelic *CEBPA* mutations;
- *RUNX1-RUNX1T1*—AML with *t(8;21)(q22;q22)*;
- *IDH2*<sup>R172</sup>—AML with *IDH2*<sup>R172</sup> mutations and no other class-defining lesions;
- *Inv(3)*—AML with *inv(3)(q21q26.2)* or *t(3;3)(q21;q26.2)*;
- Other—samples that are not assigned to any of above subgroups.

More detailed information of this genomic classification for the multiple mutations and “other” group are described in the supplementary document.

## 2.3 | Systematic identification of subtype-specific genes

A subtype-specific gene must be over-expressed in a single subtype compared to all the other subtypes while the other subtypes are not statistically different from each other. However, the problem is complicated due to the fact that we have 11 subtypes. If we just test the difference between one subtype against the rest, it only implies that this subtype is different from the combined distribution of the rest but the other subtypes could also differ from each other, such that the specificity to one single subtype cannot be guaranteed.

To overcome this issue, we apply a method that is originally described in<sup>19</sup> and use 11 AML subtypes, as described in section 2.2, for identifying subtype-specific genes. This method computes two statistics— a robust *t* test (T1) and chi-square test (T2) for each subtype. So, T1 is used to determine if there is a significant difference between each single subtype and all other subtypes. And, T2 is used to test if the other subtypes have similar expression. To be considered as subtype-specific, the statistic of T1 must be large so that the subtype is significantly differential from the others, but the statistic of T2 must be small to control the similarity between the remaining subtypes. The statistical significance is expressed in terms of false discovery rate (FDR) to account for multiple testing.<sup>20</sup> Following,<sup>19</sup> we set the threshold of T1-based FDR < 0.01 and T2-based FDR > 0.10.

## 2.4 | Code availability

The codes to compute two statistics T1 and T2 for identifying subtype-specific genes used in this study are adapted from the original study<sup>19</sup> and can be downloaded from the public Zenodo repository at <https://doi.org/10.5281/zenodo.4036552>.

## 3 | RESULTS

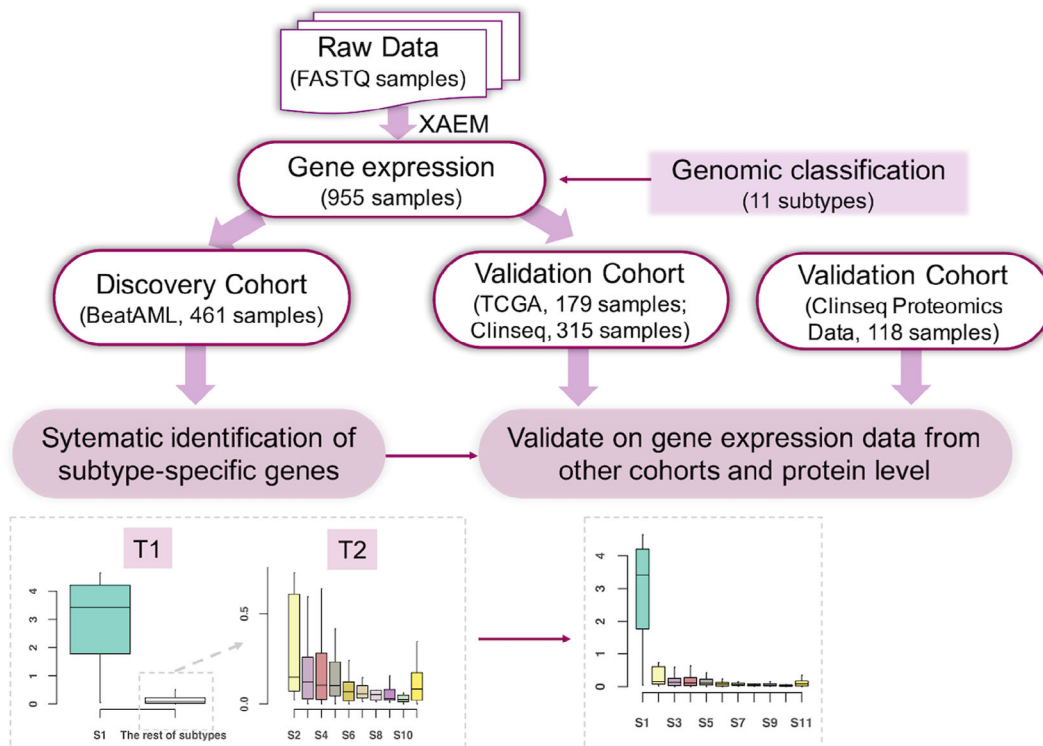
The pipeline of the subtype-specific genes discovery analysis is illustrated in Figure 1. Further experimental details are described in materials and methods section. Table S1 shows the number of patients in each subtype; *NPM1* is the largest group across three cohorts (including 22.5% of samples), which is in agreement with previous reports.<sup>6,21</sup> The *TP53*-mutant subtype is of similar proportion to the *NPM1* subtype, and the splice subtype is in the 3rd largest group. These three subtypes contain more than half of samples, accounting for 54.8% of the three cohorts. Some small subgroups, that is, *CBFB-MYH11*, *PML-RARA*, *RUNX1-RUNX1T1*, *Inv(3)* and *MLL* fusions, represent less than 5% of the three cohorts, in line with the WHO classification.

The complete result for all genes can be found at an interactive website <https://ngiavtr.shinyapps.io/AMLSubtypeSpecificDiscovery/>.

### 3.1 | Using this resource to enable the discovery of subtype-specific genes

As an example of gene-level distributions, Figure 2 presents the boxplots of the gene-level mRNA expression of *PTPRG*, a top-ranking gene specific to the *PML-RARA* subtype. It shows that *PTPRG* expression patterns are similar across the three cohorts, with high expression in the *PML-RARA* and low expression in the other subtypes, which is in agreement with other analysis.<sup>22,23</sup> The statistics of our method to identify subtype-specific genes are: T1  $\approx 0$  (*t*-statistic  $\approx 73.56$ ), T2  $\approx 1$  (Table S2). The *PTPRG* expression is also validated at the protein level with higher protein levels in *PML-RARA* compared to all the other subtypes, as shown in Figure 2D. The top five genes in each subtype are given in Table S2. To select these, we first filter out the genes with T2-based FDR < 0.1, then rank the genes by T1.

Figure S1 shows the number of subtype-specific genes assigned to each subtype based on the BeatAML as the discovery set. There were 9226 subtype-specific genes across the 11 subtypes, with the *NPM1* subtype being the largest group with 3687 genes (40%). We also investigate if the genes specific to the *NPM1* subtype are able to separate patients with and without *FLT3-ITD* (+/-) within this subtype. Figure S2 shows a tSNE plot of mRNA expressions of the top 25 subtype-specific genes for the two groups of *FLT3-ITD* +/- . The result shows no separation between two groups, indicating these subtype-specific genes are specific to *NPM1* but do not contain signals for the *FLT3-ITD* status. The subtype “other” is the smallest group including only 52 genes (0.6%),



**FIGURE 1** Pipeline of systematic identification of subtype-specific genes from AML RNA-seq, using the BeatAML data as discovery set, and the TCGA and Swedish Clinseq data sets as the validation set. The statistic T1 finds genes that are over-expressed in a single subtype compared to all the other subtypes, and T2 limits to genes where the remaining subtypes are not statistically different from each other [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

which is reasonable since this group has no particular criteria associated. The distribution of T1 of the identified subtype-specific genes for each subtype based on the BeatAML dataset are shown in Figure S3.

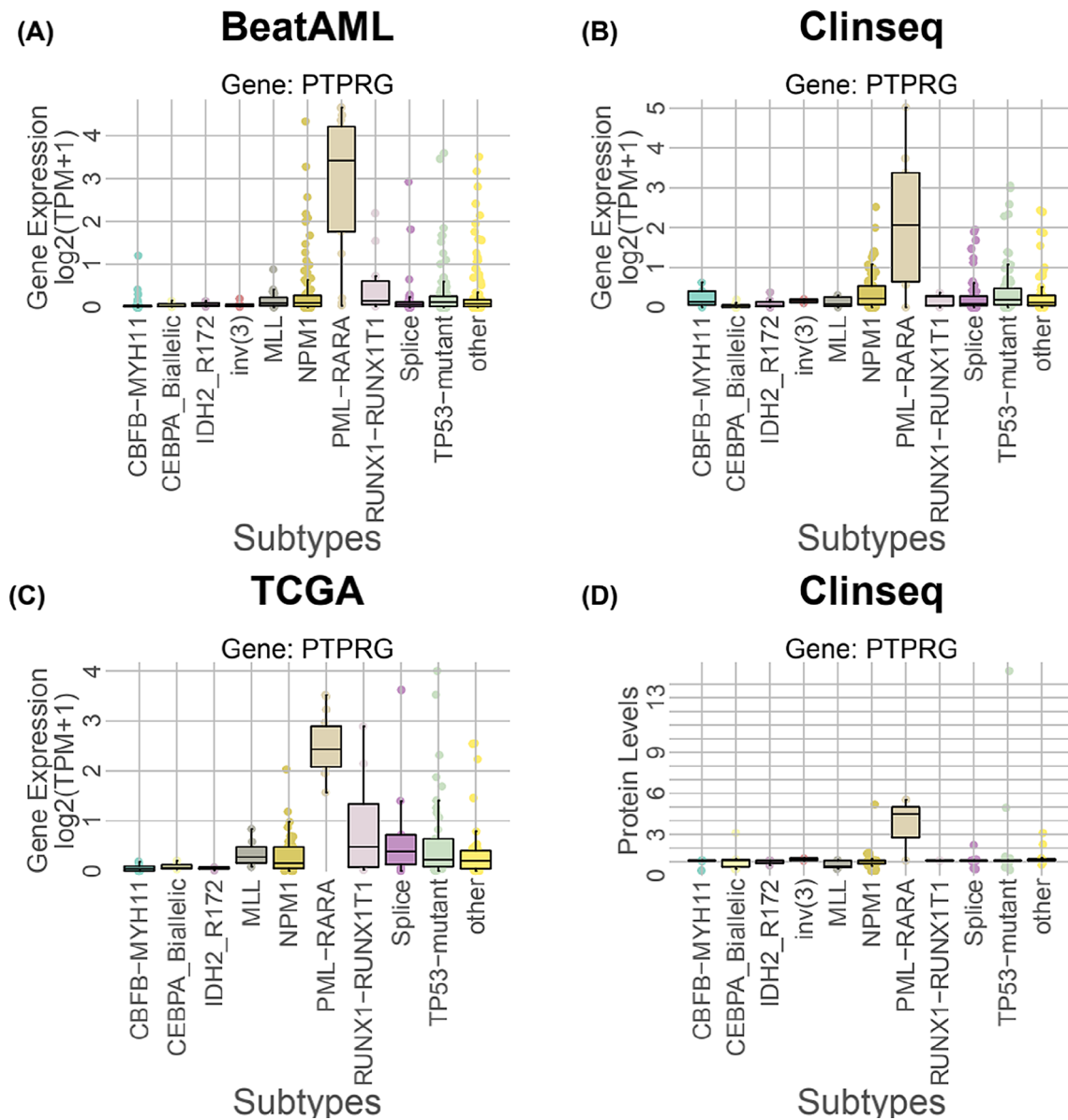
### 3.2 | Validation of subtype-specific genes

Figure S4 shows the Venn diagrams the subtype-specific genes from the three cohorts for each subtype. In total, 729 subtype-specific genes discovered in the BeatAML cohort are validated in both the TCGA and Clinseq cohorts. The *NPM1* and *PML-RARA* subtypes have significantly higher numbers of overlapping genes across the three cohorts, with 240 and 210 respectively, compared to other subtypes ( $\leq 110$ ).

In the heat map (Figure 3A) displaying the top 25 genes from each subtype in the discovery set (BeatAML), a pattern that distinguishes the different subtypes is evident. For patients with the *NPM1*, *CEBPA<sup>Biallelic</sup>*, *CBFB-MYH11*, *PML-RARA*, *MLL* and splice subtypes, these data suggest gene expression patterns specific to the corresponding subtypes. Similar patterns are observed in the validation sets (Clinseq and TCGA) for the *NPM1*, *CEBPA<sup>Biallelic</sup>*, *CBFB-MYH11* and *PML-RARA* subtypes, as shown in Figure 3B, C. Taking the *NPM1* subtype as an example, the top 25 subtype-specific genes are clearly more highly expressed in the patients classified in the *NPM1* subtype than in the other subtypes. Furthermore, this pattern

is also observed in the validation sets. Since the expression data of some genes are not provided in the TCGA data set, we indicate these absent genes with white lines (Figure 3C). Based on the clinical information from the TCGA data set, no sample belongs to *inv(3)* subtype with the Papaemmanuil et al. classification.

Furthermore, the Jaccard similarity coefficient was calculated to measure the similarity between each pair of cohorts. It is defined as the size of the intersection divided by the size of the union of the subtype-specific gene sets (Table S3), thus a higher proportion indicates a higher similarity between sample sets. In general, similar proportions are shown for different pairs of cohorts. Subtypes with good survival including *CBFB-MYH11*, *PML-RARA* and *CEBPA<sup>Biallelic</sup>* generally have higher Jaccard coefficients between the cohorts, ranging from 12% to 20%. The coefficients of poor survival subtypes such as *TP53*-mutant and *MLL* are lower. The *NPM1* subtype is also distinct from the other subtypes that its Jaccard similarity coefficient between the BeatAML and Clinseq cohorts is the highest (28%), indicating the concordance between the two cohorts. However, the coefficients for the *NPM1* group between the TCGA and other cohorts are much lower (8% and 10% for BeatAML and Clinseq, respectively). This suggests a large difference in the *NPM1* subtype between the TCGA from the BeatAML and Clinseq cohorts. In addition, to quantify the similarity of gene expression distributions of the identified subtype-specific genes between cohorts, we used the set of top 25 subtype-specific genes to calculate the correlations of median expression for each subtype between BeatAML and other cohorts (Figure S5). The result shows



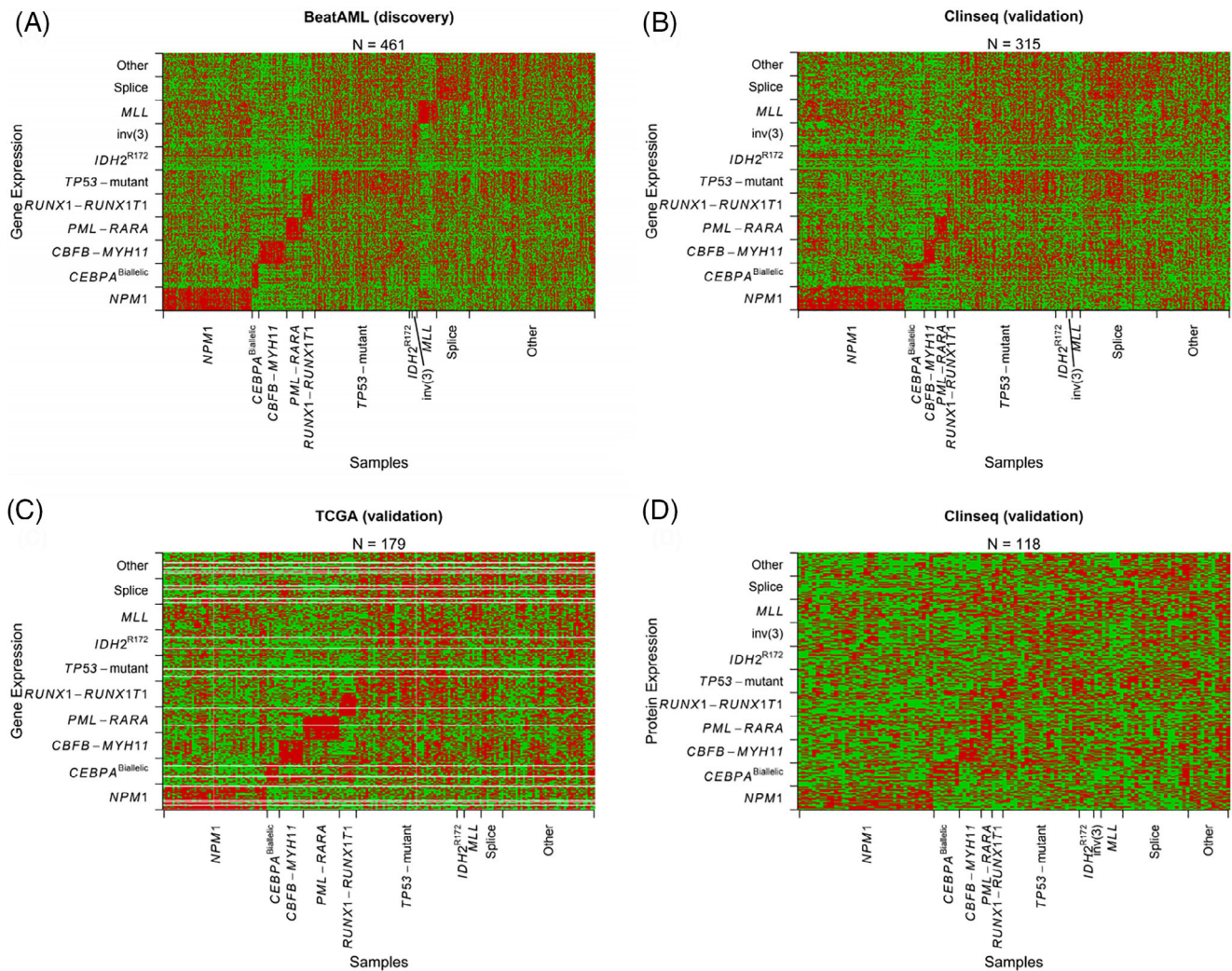
**FIGURE 2** Gene-level expression distribution of *PTPRG* gene across 11 molecular subtypes in A, BeatAML, B, Clinseq and C, TCGA cohort and protein expression distribution of *PTPRG* gene across 11 molecular subtypes in D, Clinseq cohort. Gene expression is illustrated in  $\log_2(\text{TPM} + 1)$  and the protein levels of the gene are logged relative ratios [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

that the gene expression for the top 25 subtype-specific genes has a strong correlation between BeatAML and other cohorts (Figure S5).

To present an overview of how well the global gene expression data corresponds to the genomic subtypes, we performed the tSNE analysis using: a) all genes and b) the top 25 subtype-specific genes in each subtype of BeatAML cohort. Using top 25 subtype-specific genes, we found a marked increase in the separation between subtypes (Figure 4). Using all genes, only the *PML-RARA* subtype is well separated (Figure S6). This suggests that most patients had gene expression characteristics consistent with the genomic subtype in the identified subtype-specific genes rather than all genes. We excluded the “other” group from the tSNE analysis to maintain the pure molecular distinct subtypes. The tSNE plot including the “other” group can be found in Figure S7. The tSNE analyses for the top 25 subtype-specific

genes identified in the TCGA and Clinseq cohorts are presented in Figure S8A,B, respectively. Similar to results based on the BeatAML cohort, both show a good separation between the subtypes. The color-maps for the top 25 subtype-specific genes identified in individual cohorts are presented in Figures S13–S15. We further use the 729 subtype-specific genes identified in the BeatAML and validated in both the TCGA and Clinseq cohorts from Figure 4 for the tSNE analyses. The results for the BeatAML, Clinseq and TCGA cohorts are presented in Figure S8C–E, respectively. The subtypes are well separated, especially the main subtypes such as *NPM1*, *TP53-mutant*, *CBFB-MYH11*, *PML-RARA*, *CEBPA<sup>Biallelic</sup>*, and *RUNX1-RUNX1T1*.

We further investigated the contribution of the subtype-specific genes to the separation of French-American-British (FAB) subtypes of AML,<sup>24</sup> which capture the level of maturation of the cancer cells. This



**FIGURE 3** Color-map of the top 25 subtype-specific genes from each of the 11 subtypes. A, BeatAML as the discovery set; B, Clinseq gene expression, C, TCGA gene expression, and D, Clinseq protein levels as validation sets. Red and green indicate expression levels above and below median, respectively. White lines in C, indicate genes that are not measured in the TCGA data [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

might indicate whether the subtype-defining mutations introduce a differentiation block and expansion. We selected two largest subgroups *NPM1* and *TP53*-mutant from the BeatAML cohort, collected samples with available FAB information and did the tSNE analysis using the mRNA expression of top 25 subtype-specific genes of the subtypes. For *NPM1* subgroup, as shown in Figure S9A, we see a clear separation for FAB subtypes of M1, M2, M4 and M5. However, for *TP53*-mutant subgroup, some FAB subtypes such as M2 are not well clustered; see Figure S9B. Even though the results are limited due to the few samples with available FAB subtype information, the tSNE analysis suggests that the expression of subtype-specific genes contain information for separation of FAB subtypes.

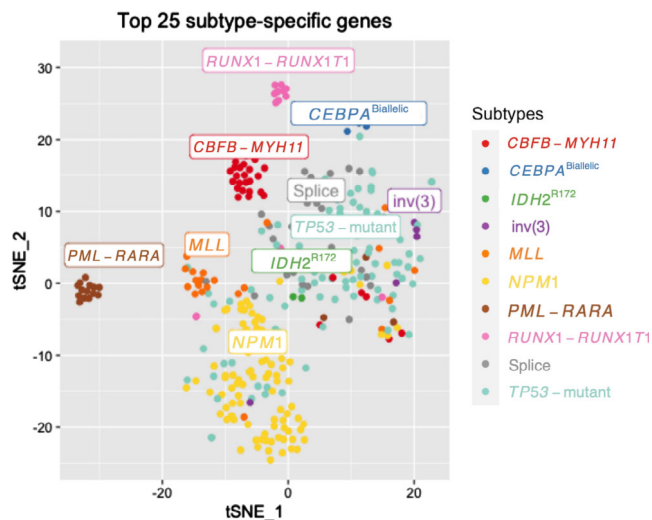
### 3.3 | Subtype-specific protein levels

The same validation process was applied to the proteomics data from the Clinseq cohort. Taking *PTPRG* gene as an example (Figure 2D), the

protein levels of *PTPRG* gene is highly correlated with gene expressions, as they all have higher values in *PML*-*RARA* subtype and lower values in the rest subtypes. Since some genes were not detected in the proteomics data, we included more genes from the subtype-specific gene list and selected the top 25 overlapping genes. So, 41% of these selected genes have high correlation between gene expression and protein levels (with correlation coefficient > 0.5). More intuitively, Figure 3 shows the protein levels of the selected genes for each subtype. Although the validation at protein level is not as clear as at the gene expression level, we still can observe consistent patterns, for example in the *NPM1* and *PML*-*RARA* subtypes.

### 3.4 | Biological pathways

To assess the relations of the identified subtype-specific genes to the biological pathways and processes, we used the Reactome database<sup>25</sup> to analyze the pathway enrichment analysis of the top 25 subtype-



**FIGURE 4** The tSNE analysis on the mRNA expression data separating genomic subtypes of AML for the top 25 subtype-specific genes in each subtype, and the data are shown for the BeatAML cohort [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

specific genes in each subtype. The significant pathways ( $p$ -value  $< .05$ ) and statistics for each subtype are listed in Table S4. For example, the most significant pathway ( $p$ -value = 0.0019) related to *CBFB-MYH11* subtype-specific genes is “*RUNX2* regulates osteoblast differentiation”. The *RUNX* family of transcription factors plays a critical role in hematopoiesis, and the *RUNX1* transcription factor is frequently translocated in AML.<sup>26</sup> As a *RUNX* family member, the relation of *RUNX2* to *CBFB-MYH11* specific genes underlines the fact that *RUNX* family members can function in complex with *CBFB*.<sup>27</sup> The roles of *RUNX2* and *CBFB* in skeletal development have also been well studied elsewhere.<sup>28,29</sup> In addition, the glance of pathway diagram would bring an overview of the connectivity and flow of information in biological systems. Taking the most significant pathway related to *CEBPA<sup>Biallelic</sup>* subtype-specific genes ( $p$ -value  $< .001$ ), “Glucuronidation”, for example, we demonstrated the pathway associated with the subtype-specific genes (Figure S10). Four of a total 24 genes of this pathway (16.7%) including *UGT2A3*, *UGT2B10*, *UGT2B11* and *UGT2B28* are represented in 25 top subtype-specific genes of *CEBPA<sup>Biallelic</sup>* subtype, indicating the pathway enrichment.

## 4 | DISCUSSION AND CONCLUSION

We have provided a comprehensive view of the transcriptomic landscape of molecular subtype-specific mRNA expression of AML based on 955 RNA-seq samples from three different cohorts. A sophisticated statistical methodology has been used to identify and validate 729 subtype-specific genes across molecular subtypes. These results suggest that the gene-expression profiles can be used to characterize the molecular subtypes of AML. We have provided a comprehensive data portal which can serve a public resource to be used for

hypothesis generation, that is, the discovery of new biomarkers for drug targets.

Rather than discovery of new genomic subtypes, our aim was to provide more characterizations to existing molecular subtypes that have been shown to have strong clinical relevance. Currently in the clinic, the WHO and the ELN classifications have the most impact on the treatment decision, but these are based on biology, morphology and medical history. We had chosen the molecular classification of Papaemmanuil et al. as it provides purely biological distinct subtypes. The value of the molecular classification will increase over time as we get more information on subtype-specific markers and on treatment response. Instead of gene signatures, we provide single gene markers specific to each subtypes. The key advantages are that (i) the single genes may have more biological investigations, and (ii) they are more easily measured in terms of sample requirements, especially if they are validated at single protein level.

The *PTPRG* gene, a top-ranking gene specific to the *PML-RARA* subtype in Figure 2, is strongly validated across cohorts and at protein level. This has been reported as a tumor-suppressor gene in not only AML but also other cancers, that is, nasopharyngeal carcinoma.<sup>30</sup> So, *PML-RARA* fusion is a known initiating event for the acute promyelocytic leukemia (APL), and *PTPRG* mutation was discovered in APL samples.<sup>31</sup> Note, *PTPRG* is a member of the protein tyrosine phosphatase (PTP) family of signaling molecules that are involved in cell cycle, differentiation and oncogenic transformation. It has been identified as a significant gene in childhood acute lymphoblastic leukemia since the phosphatase induces dephosphorylation of *ERK* which provides a potential therapeutic target for RAS-related leukemias.<sup>32</sup> Also, PTPs have been detected in AML previously,<sup>33</sup> but not with specificity to any subtypes. The current result suggests a biological role of *PTPRG* specific to the *PML-RARA* fusion subtype. Thus, by identifying subtype-specific genes we can discover biomarkers that are specific to each subtype. These genomic biomarkers could have more biological investigations by clinical researchers.

Different molecular subtypes are clearly distinguished from each other in the discovery set (BeatAML) and similar patterns are observed in the validation sets (TCGA and Clinseq), especially for *NPM1*, *CEBPA<sup>Biallelic</sup>*, *CBFB-MYH11*, *PML-RARA* and *RUNX1-RUNX1T1*. This partially agreed with the analysis in,<sup>34</sup> where they clustered patients and genes on the basis of similarity of expression distributions by using unsupervised hierarchical clustering analysis and found patients with *t(8;21)*, *inv(16)* mutations—*RUNX1-RUNX1T1* and *CBFB-MYH11* fusions—had gene expression patterns specific to *t(8;21)* and *inv(16)* subtypes. Among the subtypes, the *PML-RARA* and *NPM1* subtypes have the largest number of subtype-specific genes. Such a large proportion of subtype-specific genes is typically a reflection of a biologically distinct entity. We observe the same phenomenon, for example, if we compare estrogen-receptor (ER)-positive vs ER-negative breast cancers, or lung adenocarcinoma vs squamous-cell carcinoma.

Further explorations could be made based on the identified subtype-specific biomarkers. For example, we have found two subtype-specific genes from *CEBPA<sup>Biallelic</sup>* subtype, *ZBTB20* and *ARHGEF6*, in the set of transcription factor (TF) targets of *CEBPA*

collected from the Molecular Signatures Database (MSigDB v7.1). The expression distributions (Figures S11 and S12) show the specificity of these genes to *CEBPA*<sup>Biallelic</sup> with a higher expression of the subtype over the rest groups in the BeatAML as well as the validated cohorts. This suggests a role of the *CEBPA* mutation in the changes of expression of the TF targets.

Our study had some limitations. Because they were based on retrospective samples, the sample collection in the different cohorts was heterogeneous. For example, not all samples across the different cohorts were taken uniformly at the time of diagnosis; also, both bone marrow and peripheral-blood mononuclear cells were in use, identified using Ficoll gradient centrifugation rather than using the CD38+ marker. These effects tend to increase statistical variability, so may reduce the sensitivity in detecting the subtype-specific markers and explain some lack of replicability across the cohorts. However, they do not reduce the specificity of our results.

#### AUTHOR CONTRIBUTIONS

Trung Nghia Vu and Yudi Pawitan initiated and coordinated the study; Tian Mou performed data analysis; Tian Mou, Yudi Pawitan, Trung Nghia Vu contributed to method development and manuscript writing. Matthias Stah, Mattias Vesterlund, Rozbeh Jafari, Anna Bohlin, Albin Österroos, Ioannis Siavelis, Helena Bäckvall, Helena Bäckvall, Tom Erkers, Santeri Kiviluoto, Brinton Seashore-Ludlow, Päivi Östling, Lukas M. Orre, Olli Kallioniemi, Sören Lehmann, Janne Lehtiö also contributed to manuscript writing and the Clinseq cohort including the acquisition and processing of patient samples, and the collection of clinical data. Wenjiang Deng and Trung Nghia Vu contributed to the pre-processing of RNA sequencing data. All authors read and approved the final manuscript.

#### FUNDING INFORMATION

This work was supported by funding from the Swedish Cancer Fonden and Barncancerfonden, the Swedish Research Council (No.2017-06095 and No. 2019-01857), the Swedish Foundation for Strategic Research (SB16-0058) and Knut and Alice Wallenberg Foundation (2015.0291). The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) in Uppsala, which is partially funded by the Swedish Research Council through grant agreement no. 2018-05973. WD is partly supported by the Chinese Scholarship Council (grant NO. 201600160085).

#### CONFLICT OF INTEREST

The authors declare no conflict of interest.

#### DATA AVAILABILITY STATEMENT

The results of subtype-specific genes discovery in this study can be found at: <https://nghiaivr.shinyapps.io/AMLSubtypeSpecificDiscovery/>.

#### ORCID

Tian Mou  <https://orcid.org/0000-0001-7707-8760>

#### REFERENCES

1. Assi SA, Imperato MR, Coleman DJL, et al. Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nat Genetics*. 2019; 51(1):151-162.
2. Network Cancer Genome Atlas Research. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.
3. Lee W, Alexeyenko A, Pernemalm J, et al. Identifying and assessing interesting subgroups in a heterogeneous population. *BioMed Res Int*. 2015;2015(3):1-13.
4. Patel JP, Goonen M, Figueroa ME, et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N Engl J Med*. 2012;366(12):1079-1089.
5. Grimwade D, Hills RK, Moorman AV, et al. Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom medical research council trials. *Blood*. 2010;116(3):354-365.
6. Papaemmanuil E, Gerstung M, Lars Bullinger VI, et al. Genomic classification and prognosis in acute myeloid leukemia. *N Engl J Med*. 2016; 374(23):2209-2221.
7. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the world health organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*. 2009; 114(5):937-951.
8. James R. Cook. Chapter 25 - molecular hematopathology. In: Tubbs RR, Stoler MH, eds. *Cell and Tissue Based Molecular Pathology*. Philadelphia: Churchill Livingstone; 2009:305-324.
9. Matsuo H, Kajihara M, Tomizawa D, et al. EVI1 overexpression is a poor prognostic factor in pediatric patients with mixed lineage leukemia-AF9 rearranged acute myeloid leukemia. *Haematologica*. 2014;99(11):e225-e227.
10. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*. 2018;562(7728):526-531.
11. Wang M, Lindberg J, Klevebring D, et al. Validation of risk stratification models in acute myeloid leukemia using sequencing-based molecular profiling. *Leukemia*. 2017;31(10):2029-2036.
12. Deng W, Mou T, Kalari KR, et al. Alternating EM algorithm for a bilinear model in isoform quantification from RNA-seq data. *Bioinformatics*. 2019;36(3):1-8.
13. Branca RMM, Lukas M Orre, Henrik J Johansson, Viktor Granholm, Mikael Huss, sa Perez-Bercoff, Jenny Forshed, Lukas Kl, and Janne Lehtiö. Hirief lc-ms enables deep proteome coverage and unbiased proteogenomics. *Nat Methods*. 2014;11(1):59-62.
14. Johansson HJ, Socciarelli F, Vacanti NM, Mads H. Haugen, and Janne Lehtiö. Breast cancer quantitative proteome and proteogenomic landscape. *Nat Commun*. 2019;10(1):1600.
15. Yafeng Zhu LM, Orre HJ, Johansson MH, et al. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat Commun*. 2018;9(1):903.
16. Moggridge S, Sorensen P, Morin GB, Hughes CS. Extending the compatibility of the SP3 paramagnetic bead processing approach for proteomics. *J Proteome Res*. 2018;17:1730-1740.
17. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017;35(4):316-319.
18. Jorrit Boekel, *GitHub*, <https://github.com/lehtiolab/ddamsproteomics/releases/tag/v1.0.2>, 2019.
19. Vu TN, Pramana S, Calza S, Suo C, Lee D, Pawitan Y. Comprehensive landscape of subtype-specific coding and non-coding RNA transcripts in breast cancer. *Oncotarget*. 2016;7(42):68851-68863.
20. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*. 2005;21(13):3017-3024.



21. Rose D, Haferlach T, Schnittger S, Perglerova K, Kern W, Haferlach C. Subtypespecific patterns of molecular mutations in acute myeloid leukemia. *Leukemia*. 2017;31(1):11-17.
22. Payton JE, Grieselhuber NR, Chang L-W, et al. High throughput digital quantification of mRNA abundance in primary human acute myeloid leukemia samples. *J Clin Invest*. 2009;119(6):1714-1726.
23. Payton JE, Grieselhuber NR, Chang L-W, et al. The expression signature of M3 AML suggests direct and indirect effects of PML-RARA on target genes. *Blood*. 2007;110(11):719.
24. Bennett JM, Catovsky D, Daniel M-T, et al. Proposals for the classification of the acute leukaemias French-American-British (FAB) co-operative group. *Br J Haematol*. 1976;33(4):451-458.
25. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498-D503.
26. Ichikawa M, Yoshimi A, Nakagawa M, Nishimoto N, Watanabe-Okochi N, Kurokawa M. A role for RUNX1 in hematopoiesis and myeloid leukemia. *Int J Hematol*. 2013;97(6):726-734.
27. Otto F, Kanegane H, Mundlos S. Mutations in the RUNX2 gene in patients with cleidocranial dysplasia. *Hum Mutat*. 2002;19(3):209-216.
28. Kundu M, Javed A, Jeon J-P, et al. Cbfb interacts with Runx2 and has a critical role in bone development. *Nat Genetics*. 2002;32(4):639-644.
29. Yoshida CA, Furuichi T, Fujita T, et al. Core-binding factor  $\beta$  interacts with Runx2 and is required for skeletal development. *Nat Genetics*. 2002;32(4):633-638.
30. Cheung AK, Ip JC, Chu AC, et al. PTPRG suppresses tumor growth and invasion via inhibition of Akt signaling in nasopharyngeal carcinoma. *Oncotarget*. 2015;6(15):13434-13 447.
31. Welch JS, Ley TJ, Link DC, et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell*. 2012;150(2):264-278.
32. Xiao J, Lee S-T, Xiao Y, et al. PTPRG inhibition by DNA methylation and cooperation with RAS gene activation in childhood acute lymphoblastic leukemia. *Int J Cancer*. 2014;135(5):1101-1109.
33. Arora D, Kothe S, Van Den Eijnden M, et al. Expression of protein-tyrosine phosphatases in acute myeloid leukemia cells: FLT3 ITD sustains high levels of DUSP6 expression. *Cell Commun Signal*. 2012;10(19):1-15.
34. Yagi T, Morimoto A, Eguchi M, et al. Identification of a gene expression signature associated with pediatric aml prognosis. *Blood*. 2003;102(5):1849-1856.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Mou T, Pawitan Y, Stahl M, et al. The transcriptome-wide landscape of molecular subtype-specific mRNA expression profiles in acute myeloid leukemia. *Am J Hematol*. 2021;96:580-588. <https://doi.org/10.1002/ajh.26141>