# An extracellular matrix signature in leukemia precursor cells and acute myeloid leukemia

*Valerio Izzi,*[1] *Juho Lakkala,*[1] *Raman Devarajan,*[1] *Heli Ruotsalainen,*[1] *Eeva-Riitta Savolainen,*[2,3] *Pirjo Koistinen,*[3] *Ritva Heljasvaara*[1,4] *and Taina Pihlajaniemi*[1]

*[1]Centre of Excellence in Cell-Extracellular Matrix Research and Biocenter Oulu, Faculty of Biochemistry and Molecular Medicine, University of Oulu, Finland; [2]Nordlab Oulu and Institute of Diagnostics, Department of Clinical Chemistry, Oulu University Hospital, Finland; [3]Medical Research Center Oulu, Institute of Clinical Medicine, Oulu University Hospital, Finland and [4]Centre for Cancer Biomarkers (CCBIO), Department of Biomedicine, University of Bergen, Norway*

*Correspondence: valerio.izzi@oulu.fi*

*doi:10.3324/haematol.2017.167304*

**Izzi et al. Supplementary Information**

Supplementary information to this submission contain **Supplementary Materials and Methods**, four Supplementary Figures (**Supplementary Fig S1-S4**) and four Supplementary Tables (**Supplementary Table S1-S4**).

## Supplementary Materials and Methods

### *Compilation of the ECM gene set*

We used gene ontology (GO) annotations from the gene ontology consortium (*http://geneontology.org/)* to define ECM genes. To this aim, we compiled an initial redundant set of 3170 genes by appending all the genes belonging to the following GO categories: GO:0005578 (proteinaceous extracellular matrix), GO:0044420 (extracellular matrix component), GO:0085029 (extracellular matrix assembly), GO:0030198 (extracellular matrix organization), GO:0005201 (extracellular matrix structural constituent), GO:0031012 (extracellular matrix), GO:0022617 (extracellular matrix disassembly), GO:0035426 (extracellular matrix-cell signaling), GO:1990430 (extracellular matrix protein binding) and GO:0070278 (extracellular matrix constituent secretion). Duplicates were manually removed and the list of candidate ECM genes was checked against The Matrisome Database (http://matrisomeproject.mit.edu/proteins/) to ensure that each gene with "ECM" GO annotations was known to produce an ECM protein. For genes not matching with The Matrisome Database (named "Not available" in the text), inclusion was decided upon screening the annotations in two additional databases, UniProt (http://www.uniprot.org/) and PSORT II (http://psort.hgc.jp/). Genes with "ECM" annotation in the two databases were appended to the final list, which contained 135 genes. This final list of ECM genes was finally subjected to the Affymetrix conversion tool (http://www.affymetrix.com/) to obtain the gene identifiers for the Human Genome U133 Plus 2.0 Array (hgu133plus2) chip type.

### *Support Vector Machine (SVM)*

The analysis of predictors (genes) that, among the whole ECM signature, best discriminated AML patients from healthy donors in each cohort was performed using IBM SPSS Modeler 18. To this aim, we used a leave-one-out, in-line validation scheme: first, each AML cohort (GSE10358 and TCGA_LAML) was standardized to the same median, split into training and test sets (80% and 20% of each cohort, respectively) and subjected to the auto-classifier analysis. The auto-classifier node included different regression, decision tree and machine-learning algorithms, among which the best performer resulted to be SVM with radial bias function (RBF) kernel, set as follows: stopping criteria 1.0E-3, regularization parameter 10, epsilon 0.1, RBF gamma 1.0, Gamma 0.1, Bias 0.0 and 3 degrees for the function. Next, the prediction from each cohort was cross-validated by applying it to the other cohort used entirely as validation set. Finally, the two cohorts were merged, randomly shuffled and subjected to the same classifier as previously trained but with feature selection activated. The accuracy of the SVM algorithm thus set was > 99% in the single cohorts and > 97.0% in the merged cohort. The feature selection step in the merged cohort resulted in a ranked list of predictors, the top-15 of which were further validated *in vitro* by quantitative real-time PCR (qRT-PCR).

### *Quantitative Real Time PCR*

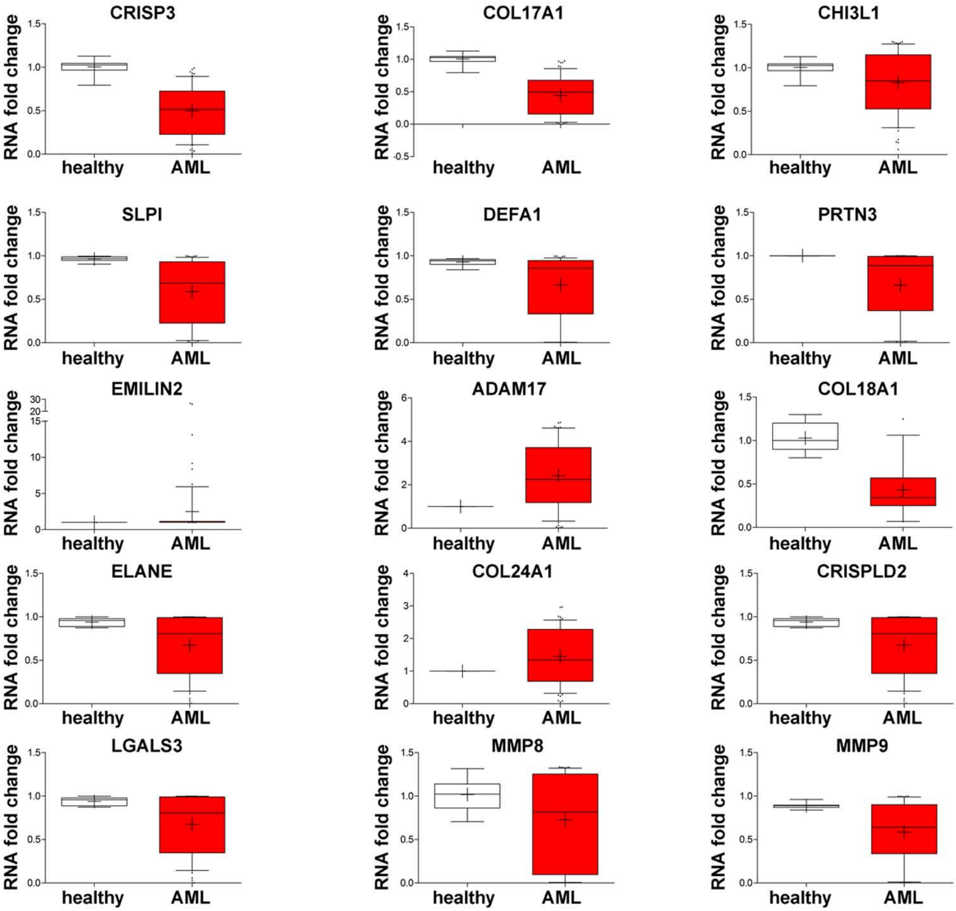
All analyses were performed on peripheral blood mononuclear cells (PBMCs) from AML samples and healthy donors. RNA was isolated with Qiagen kits (QIAzol and RNeasy mini), and cDNA was produced with the iScript cDNA synthesis kit (Bio-Rad). RT-qPCR was performed with iTaq SYBR Green Supermix with ROX reagents (Bio-Rad). All assays were performed in duplicate using a

CFX96 Real-Time System (Bio-Rad). Values in all samples were normalized to GAPDH, and fold-change ($2^{\Delta\Delta}$Cq) was calculated using CFX Manager software (Bio-Rad). The following primes were used:

| | |
|---|---|
| ADAM17 fwd (5’ -> 3’) | GTAAAACGACGGCCAGTACACCTGATAGACCCAGCTCC |
| ADAM17 rev (3’ -> 5’) | TGGCGGTAGAATCTTCCCAG |
| CHI3L1 fwd | TGAGGCATCGCAATGTAAG |
| CHI3L1 rev | AAGGGGAAGTAGGATAGGGG |
| COL17A1 fwd | TTGTTTAAGCCACCCAGTCC |
| COL17A1 rev | CAGGGGGCCTAAAGACTAGC |
| COL18A1 fwd | TGCCCATCGTCAACCTCAAG |
| COL18A1 rev | CAGAGCCTGAGAACAGAGCC |
| COL24A1 Fwd | GATTACCTGGTCATGTGGGGG |
| COL24A1 Rev | ACATCACCTTGCAGTCCTCG |
| CRISP3 fwd | TGAAGATTGATCTAGTAGCTTGCC |
| CRISP3 rev | GTAAAACGACGGCCAGTTCTGCTGGCTCCATGTGAC |
| CRISPLD2 Fwd | CTCAGCAAATACAAACCTTCCA |
| CRISPLD2 Rev | GGTCGTGTAGCAGTCCAA |
| DEFA1 fwd | GACTGCTGTCTGCCCTCTCT |
| DEFA1 rev | TTTGGGATGAGGAAAGGAAA |
| ELANE fwd | CGTGGCGAATGTAAACGTCC |
| ELANE rev | TTTTCGAAGATGCGCTGCAC |
| EMILIN 2 fwd | GTGAACATGGCCACTGACTT |
| EMILIN 2 rev | GCCCTCAGAGTGTAGATACAG |
| GAPDH fwd | GCATGGCCTTCCGTGTTC |
| GAPDH rev | CCTGCTTCACCACCTTCTTGAT |
| LGALS3 Fwd | TGCTGATAACAATTCTGGGCAC |
| LGALS3 Rev | TGAAGCGTGGGTTAAAGTGGA |
| MMP8 Fwd | AAAAGCATATCAGGTGCCTTTCCA |
| MMP8 Rev | CAGCCACATTTGATTTTGCTTCAG |
| MMP9 Fwd | GGGACGCAGACATCGTCATC |
| MMP9 Rev | TCGTCATCGTCGAAATGGGC |

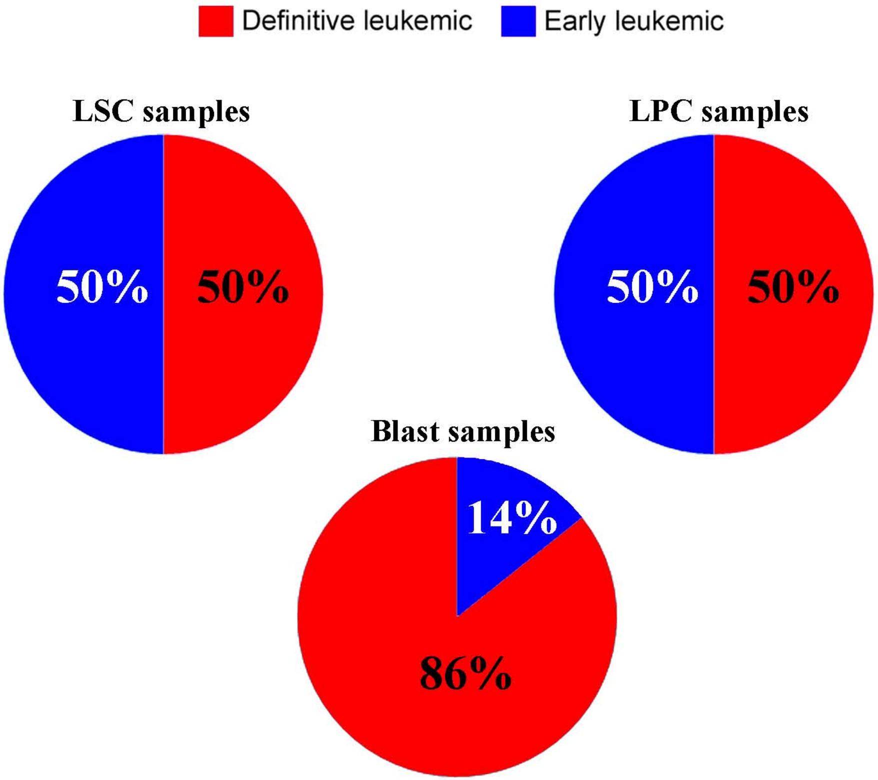
| | |
|---|---|
| PRTN3 fwd | TCTGCCGGCCACATAACATT |
| PRTN3 rev | AGAAGTCAGGGAAAAGGCGG |
| SLPI fwd | GGGAGGTCTCCCGAAACTAAG |
| SLPI rev | GTAAAACGACGGCCAGTGCAATAGTAGCTGGGAGAGGC |

**Izzi et al Supplementary Figure S1.**



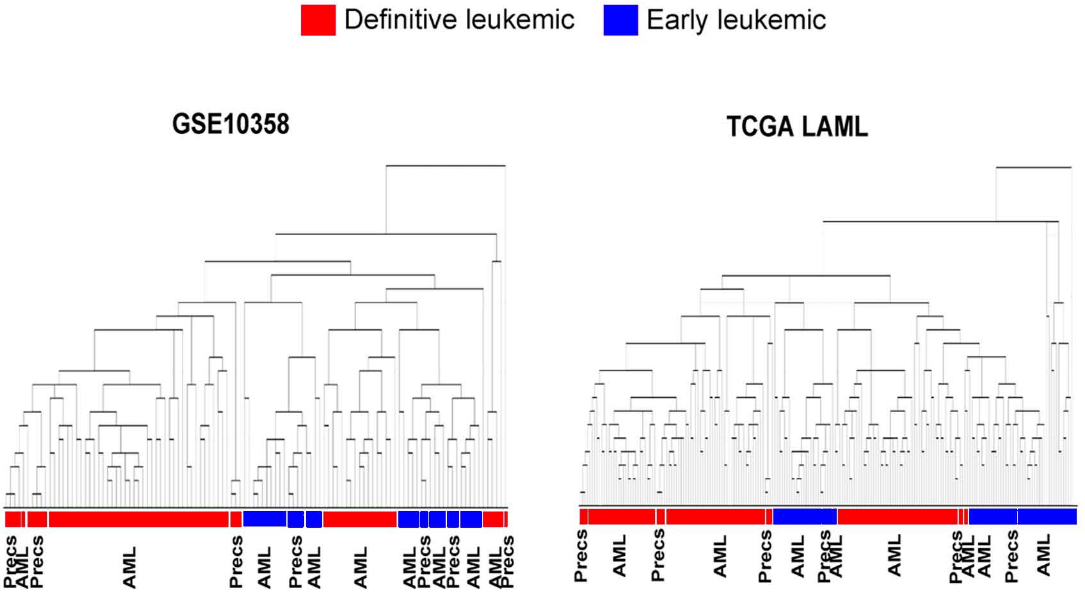
**Supplementary Figure S1. Validation of the ECM signature.** Quantitative real-time PCR (qRT-PCR) of the 15 gene with the highest predictive potential as determined by SVM analysis (see Supplementray Information) in peripheral blood mononuclear cells (PBMCs) from AML patients and healthy donors Data are reported as 10-90 percentile with outliers, median (thin internal line), mean (thin internal cross) and standard deviation. All data reported in the figure are significant at $P < 0.0001$, as from Mann-Whitney U test.

**Izzi et al Supplementary Figure S2.**



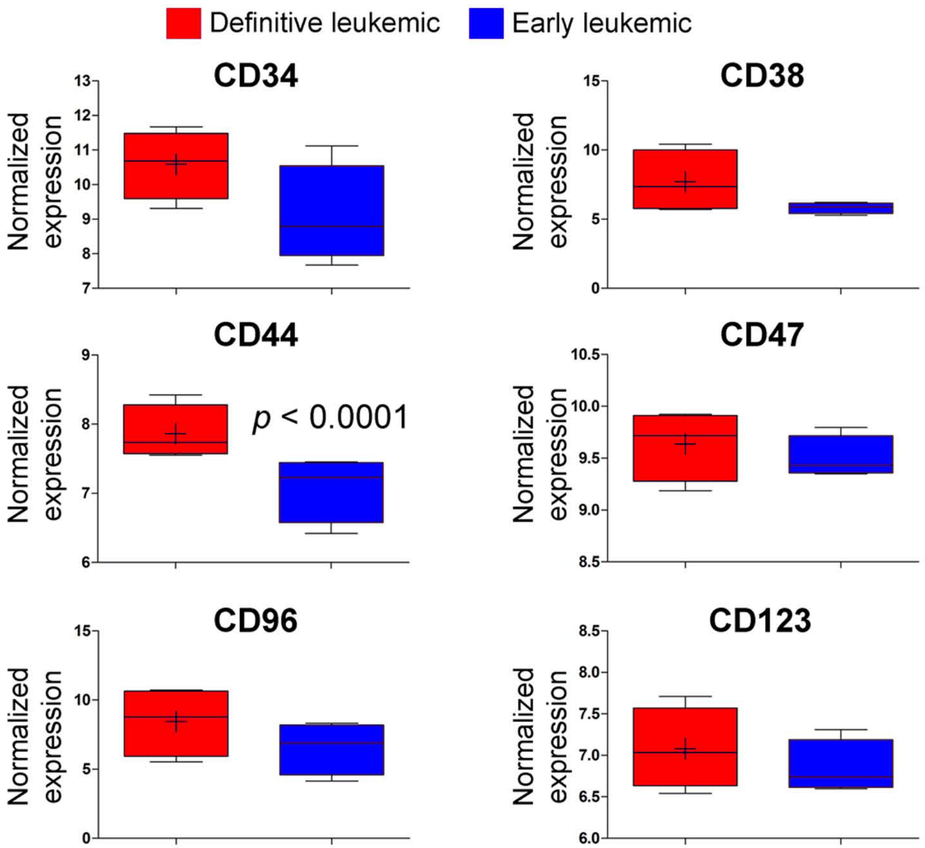
**Supplementary Figure S2. Precursors burden in early and definitive leukemic groups.** Amount of leukemia stem cells (LSC), leukemia precursor cells (LPC) and blasts in the the two groups, as assessed by the expression of the ECM signature genes.

**Izzi et al Supplementary Figure S3.**



**Supplementary Figure S3. Clustering of AML patients into early and definitive leukemic type groups.** The expression of the ECM genes was used to cluster patients from the TCGA LAML and GSE10358 cohorts into early- and definitive-type subgroups, by subjectim them to hierarchical clustering (Ward's method) together with LSC and LPC previously identified as early and definitive leukemic.

**Izzi et al Supplementary Figure S4.**



**Supplementary Figure S4. Expression of surface markers in early and definitive LSCs.** Normalized expression of the principal surface markers of LSCs in the early and defintive subgroups. To rule out possible quantitative differences in sample compostition, only LSCs in the two groups (4 each) were analyzed. Data are reported as 10-90 percentile with outliers, median (thin internal line), mean (thin internal cross) and standard deviation. *P* value is from Mann-Whitney U test.

**Supplementary Table 1. An ECM signature in leukemia precursor cells and acute myeloid leukemia.**

| Gene symbol | Ensembl gene ID | Description | Average Z fold change |
|---|---|---|---|
| TPSAB1 | ENSG00000172236 | tryptase alpha/beta 1 | -0,798 |
| EMILIN2 | ENSG00000132205 | elastin microfibril interfacer 2 | -0,228 |
| MAMDC2 | ENSG00000165072 | MAM domain containing 2 | 0,863 |
| ADAM17 | ENSG00000151694 | ADAM metallopeptidase domain 17 | -0,992 |
| COL4A5 | ENSG00000188153 | collagen type IV alpha 5 chain | -0,393 |
| COL24A1 | ENSG00000171502 | collagen type XXIV alpha 1 | -0,130 |
| MMP2 | ENSG00000087245 | matrix metallopeptidase 2 | 0,620 |
| AGRN | ENSG00000188157 | agrin | 0,371 |
| BMP1 | ENSG00000168487 | bone morphogenetic protein 1 | -0,198 |
| ADAMTSL4 | ENSG00000143382 | ADAMTS like 4 | 1,156 |
| ECM1 | ENSG00000143369 | extracellular matrix protein 1 | 0,331 |
| TIMP1 | ENSG00000102265 | TIMP metallopeptidase inhibitor 1 | -0,414 |
| OLFML2A | ENSG00000185585 | olfactomedin like 2A | -0,174 |
| IGFBP7 | ENSG00000163453 | insulin like growth factor binding protein 7 | -0,868 |
| ADAMTS2 | ENSG00000087116 | ADAM metallopeptidase with thrombospondin type 1 motif 2 | -1,841 |
| LAMB2 | ENSG00000172037 | laminin subunit beta 2 | 0,399 |
| SPON1 | ENSG00000262655 | spondin 1 | -1,022 |
| P4HA1 | ENSG00000122884 | prolyl 4-hydroxylase subunit alpha 1 | -1,099 |
| CST3 | ENSG00000101439 | cystatin C | -5,301 |
| MMP19 | ENSG00000123342 | matrix metallopeptidase 19 | -2,491 |
| CRTAP | ENSG00000170275 | cartilage associated protein | -1,294 |
| PAPLN | ENSG00000100767 | papilin, proteoglycan like sulfated glycoprotein | -0,377 |
| ADAMTS6 | ENSG00000049192 | ADAM metallopeptidase with thrombospondin type 1 motif 6 | 1,958 |
| FGFBP3 | ENSG00000174721 | fibroblast growth factor binding protein 3 | 2,131 |
| COL9A2 | ENSG00000049089 | collagen type IX alpha 2 | 0,253 |
| HAPLN4 | ENSG00000187664 | hyaluronan and proteoglycan link protein 4 | -1,064 |
| CPA6 | ENSG00000165078 | carboxypeptidase A6 | 0,214 |
| OGN | ENSG00000106809 | osteoglycin | -7,064 |
| MMP28 | ENSG00000271447 | matrix metallopeptidase 28 | -1,669 |
| ADAMTS19 | ENSG00000145808 | ADAM metallopeptidase with thrombospondin type 1 motif 19 | 0,429 |
| ANXA2 | ENSG00000182718 | annexin A2 | 0,470 |
| ADAMTSL2 | ENSG00000197859 | ADAMTS like 2 | -2,856 |
| ADAM11 | ENSG00000073670 | ADAM metallopeptidase domain 11 | -1,249 |
| COL1A1 | ENSG00000108821 | collagen type I alpha 1 | -1,047 |
| KAZALD1 | ENSG00000107821 | Kazal type serine peptidase inhibitor domain 1 | -5,242 |
| MUC5B | ENSG00000117983 | mucin 5B, oligomeric mucus/gel-forming | 0,801 |
| SFRP1 | ENSG00000104332 | secreted frizzled related protein 1 | -3,169 |
| SDC3 | ENSG00000162512 | syndecan 3 | 0,667 |
| IGF1 | ENSG00000017427 | insulin like growth factor 1 | -0,809 |
| ADAMTS13 | ENSG00000160323 | ADAM metallopeptidase with thrombospondin type 1 motif 13 | -1,109 |

| | | | |
|---|---|---|---|
| WNT5B | ENSG00000111186 | Wnt family member 5B | 0,338 |
| ANG | ENSG00000214274 | angiogenin | -0,783 |
| SDC1 | ENSG00000115884 | syndecan 1 | 0,266 |
| MMP27 | ENSG00000137675 | matrix metallopeptidase 27 | -0,350 |
| GFOD2 | ENSG00000141098 | glucose-fructose oxidoreductase domain containing 2 | 0,688 |
| FBLN5 | ENSG00000140092 | fibulin 5 | -0,248 |
| ADAM8 | ENSG00000151651 | ADAM metallopeptidase domain 8 | 0,605 |
| BMP2 | ENSG00000125845 | bone morphogenetic protein 2 | -2,153 |
| P4HB | ENSG00000185624 | prolyl 4-hydroxylase subunit beta | 2,793 |
| ACHE | ENSG00000087085 | acetylcholinesterase (Cartwright blood group) | -1,734 |
| CTSS | ENSG00000163131 | cathepsin S | 0,481 |
| COL9A3 | ENSG00000092758 | collagen type IX alpha 3 | 1,837 |
| CFP | ENSG00000126759 | complement factor properdin | -1,880 |
| FBN2 | ENSG00000138829 | fibrillin 2 | -0,619 |
| THBS4 | ENSG00000113296 | thrombospondin 4 | -0,128 |
| SULF2 | ENSG00000196562 | sulfatase 2 | -5,283 |
| VEGFA | ENSG00000112715 | vascular endothelial growth factor A | -5,251 |
| CTSL | ENSG00000135047 | cathepsin L | -0,281 |
| VCAN | ENSG00000038427 | versican | 0,218 |
| APOE | ENSG00000130203 | apolipoprotein E | 0,710 |
| CRISPLD2 | ENSG00000103196 | cysteine rich secretory protein LCCL domain containing 2 | 0,496 |
| TNXA | ENSG00000248290 | tenascin XA (pseudogene) | -0,973 |
| MMP25 | ENSG00000008516 | matrix metallopeptidase 25 | 0,442 |
| COL18A1 | ENSG00000182871 | collagen type XVIII alpha 1 chain | -3,597 |
| LGALS3 | ENSG00000131981 | galectin 3 | -0,493 |
| TFF3 | ENSG00000160180 | trefoil factor 3 | -0,325 |
| COL17A1 | ENSG00000065618 | collagen type XVII alpha 1 | -1,831 |
| TGFBI | ENSG00000120708 | transforming growth factor beta induced | -0,465 |
| CTSG | ENSG00000100448 | cathepsin G | -3,608 |
| MEGF9 | ENSG00000106780 | multiple EGF like domains 9 | 0,556 |
| SERPINA1 | ENSG00000197249 | serpin family A member 1 | -1,130 |
| ELANE | ENSG00000197561 | elastase, neutrophil expressed | -2,356 |
| APP | ENSG00000142192 | amyloid beta precursor protein | -2,516 |
| SLPI | ENSG00000124107 | secretory leukocyte peptidase inhibitor | -1,103 |
| PRTN3 | ENSG00000196415 | proteinase 3 | 0,690 |
| DEFA1 | ENSG00000206047 | defensin alpha 1 | -0,915 |
| MMP9 | ENSG00000100985 | matrix metallopeptidase 9 | -1,413 |
| CRISP3 | ENSG00000096006 | cysteine rich secretory protein 3 | -1,338 |
| MMP8 | ENSG00000118113 | matrix metallopeptidase 8 | -1,146 |
| CHI3L1 | ENSG00000133048 | chitinase 3 like 1 | -0,394 |

Data are presented as the average of the standardized fold change (average Z fold change, log scale) of each gene in each AML cohorts or in AML precursors in respect to the same gene in the healthy donors or in the healthy precursors.

**Supplementary Table 2. The Oulu AML retrospective cohort**

**Total individuals = 61**

| | |
|---|---|
| males | 30 |
| females | 31 |

**FAB classification**

| | |
|---|---|
| M0 | 5 |
| M1 | 8 |
| M2 | 20 |
| M3 | 3 |
| M4 | 19 |
| M5a | 5 |
| M5b | 5 |

**Cytogenetic findings**

| | |
|---|---|
| AML with karyotypical abnormalities | 23 |
| AML with normal karyotype | 42 |

**Supplementary Table 3. Fisher's Exact (2-sided) tests for dependencies of the early and definitive leukemic groups**

| Factor tested | GSE103580 | TCGA LAML |
|---|---|---|
| Age | 0.255 | 0.401 |
| Gender | 0.15 | 0.47 |
| Ethnicity | 0.429 | 0.078 |
| FAB | 0.08 | 0.121 |
| Karyotype | 0.102 | 0.585 |
| Molecular Abnormalities | 0.079 | 0.18 |
| Risk (cytological) | - | 0.138 |
| Risk (molecular) | - | 0.206 |

**Supplementary Table 4. Analysis of the CD44 network.**

| **Topological data** | | | |
|---|---|---|---|
| Number of nodes | 7 | | |
| Number of edges | 7 | | |
| Average node degree | 2 | | |
| Clustering coefficient | 0.786 | | |
| PPI enrichment *P*-value | 1.28e-07 | | |
| | | | |
| **Enrichment analysis*** | | | |
| **Biological processes (GO)** | | | |
| Pathway ID | Description | Count in gene set | FDR |
| GO:0022617 | ECM disassembly | 6 | 2.26e-09 |
| GO:0030574 | Collagen catabolic process | 3 | 0.00367 |
| GO:0040011 | Locomotion | 5 | 0.0113 |
| GO:0006928 | Movement of cell | 5 | 0.0172 |
| GO:0002446 | Neutrophil mediated immunity | 2 | 0.0197 |
| | | | |
| **Molecular function (GO)** | | | |
| GO:0004175 | Endopeptidase activity | 4 | 0.0101 |
| GO:0004222 | Metalloendopeptidase activity | 3 | 0.0101 |
| | | | |
| **Cellular component (GO)** | | | |
| GO:0005605 | Basal lamina | 2 | 0.0378 |
| | | | |
| **KEGG pathways** | | | |
| 04512 | ECM-receptor interaction | 3 | 0.000632 |
| 05146 | Amoebiasis | 3 | 0.000632 |
| 05200 | Pathways in cancer | 3 | 0.0121 |
| 05222 | Small cell lung cancer | 2 | 0.0248 |
| | | | |
| **INTERPRO protein domains and features** | | | |
| IPR024079 | Metallopeptidase, catalytic domain | 2 | 0.0179 |

* Enrichment was calculated using the whole genome as the statistical background. PPI: protein-protein interaction; FDR: false discovery rate.