Supplementary Information Oatley, Vargel-Bölükbası et al.

Single-cell transcriptomics identifies CD44 as a new marker and regulator of haematopoietic stem cells development

Morgan Oatley[1*], Özge Vargel-Bölükbası[1,2*], Valentine Svensson[3,4,5], Maya Shvartsman[1], Kerstin Ganter[1], Katharina Zirngibl[6], Polina V. Pavlovich[1,7], Vladislava Milchevskaya[6,8], Vladimira Foteva[1], Kedar N. Natarajan[3,9], Bianka Baying[10], Vladimir Benes[10], Kiran R. Patil[6], Sarah A. Teichmann[3] & Christophe Lancrin[1,11]

[1] European Molecular Biology Laboratory, EMBL Rome - Epigenetics and Neurobiology Unit, Monterotondo, Italy.

[2] Current address: Boston's Children Hospital/Harvard Medical School, Boston, USA.

[3] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, UK.

[4] European Molecular Biology Laboratory, EMBL-EBI, Wellcome Genome Campus, Hinxton, UK.

[5] Current address: Pachter Lab, Caltech, California, USA.

[6] European Molecular Biology Laboratory, Structural and Computational Biology Unit, Heidelberg, Germany

[7] Moscow Institute of Physics and Technology, Institutskii Per. 9, Moscow Region, Dolgoprudny 141700, Russia.

[8] Current address: Institut für Medizinische Statistik und Bioinformatik, Köln, Germany

[9] Current address: The University of Southern Denmark, Danish Institute for Advanced Study, Department of Biochemistry and Molecular Biology, Odense, Denmark.

[10] European Molecular Biology Laboratory, Genomics Core Facility, Heidelberg, Germany.

[11] Correspondence: christophe.lancrin@embl.it

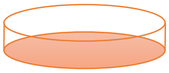
[*] Co-first authors

Supplementary Figure S1: Experimental layout for the experiments for antibody screen and single-cell RNA sequencing

a

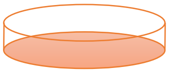
1) Differentiation of ESCs into blood cells

ESCs



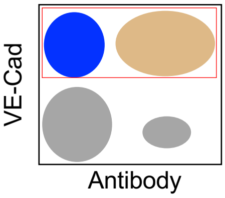
EB culture for 3 days

2) Isolation of Flk-1+ BL-CFCs

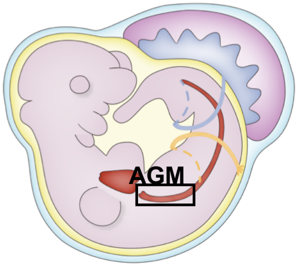


Haemangioblast culture for 1.5 days

3) FACS Analysis for VE-cadherin, CD41 and a panel of 176 markers (BD mouse Lyoplate)



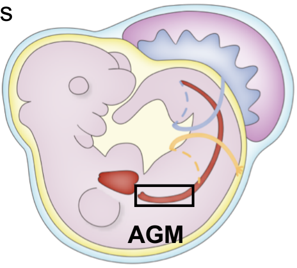
4) Identification of forty-two markers expressed by VE-Cad+. Sixteen of them displayed bimodal expression.



5) Isolation of the AGM region and test the expression of the 16 markers

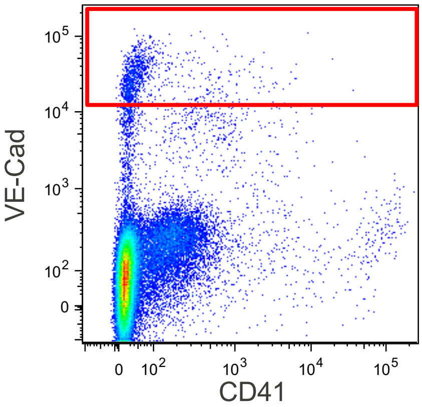
b

1) Isolation of the AGM region and analysis of VE-Cadherin and CD41 markers



2) FACS sorting of VE-Cadherin+ cells

AGM E10.5



3) Capture of 96 cells on the Fluidigm C1 platform

4) Preparation of 96 cDNA libraries and next generation sequencing on Illumina HiSeq.

5) Sequencing analysis, identification of subpopulations and selection of candidate marker genes.

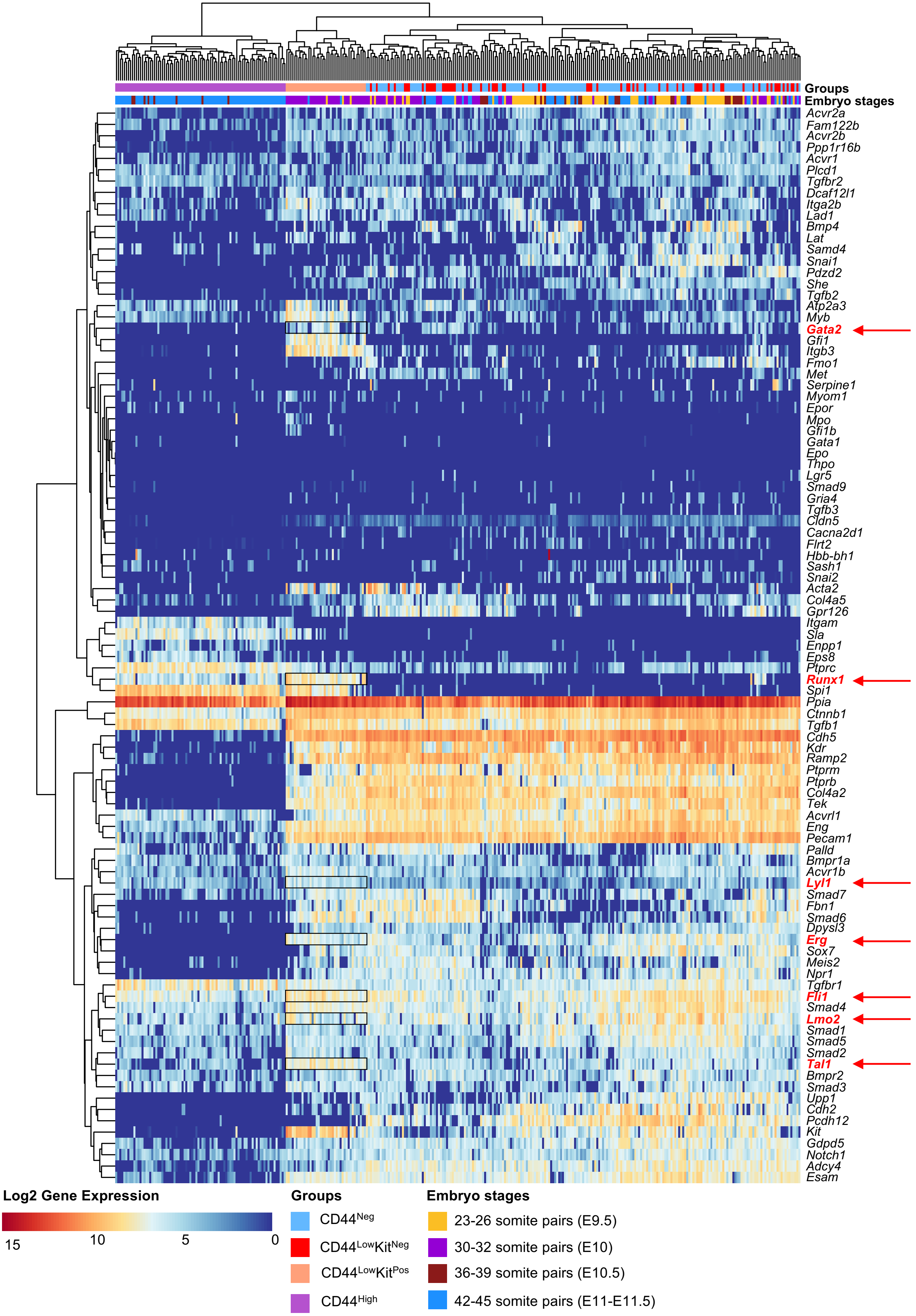
Supplementary Figure S1: Experimental layout for the experiments for antibody screen and single-cell RNA sequencing

(**a**) Strategy used for the antibody screen. (**b**) Description of the different steps following for the single cell RNA sequencing analysis in the AGM. See also Table S1 and Figure 1.

Supplementary Table S1: Results of the antibody screen

Target protein	**Gene symbol**	**Gene name**	**Bimodal expression**
CD9	Cd9	CD9 antigen	No
CD13	Anpep	alanyl (membrane) aminopeptidase	No
CD19	Cd19	CD19 antigen	No
CD23	**Fcer2a**	**Fc receptor, IgE, low affinity II, alpha polypeptide**	**Yes**
CD24	Cd24a	CD24a antigen	No
CD29	Itgb1	integrin beta 1 (fibronectin receptor beta)	No
CD31	Pecam1	platelet/endothelial cell adhesion molecule 1	No
CD34	**Cd34**	**CD34 antigen**	**Yes**
CD35	Cr2	complement receptor 2	No
CD38	Cd38	CD38 antigen	No
CD41	**Itga2b**	**integrin alpha 2b**	**Yes**
CD44	**Cd44**	**CD44 antigen**	**Yes**
CD47	Cd47	CD47 antigen	No
CD49d	**Itga4**	**integrin alpha 4**	**Yes**
CD49e	Itga5	integrin alpha 5 (fibronectin receptor alpha)	No
CD51	**Itgav**	**integrin alpha V**	**Yes**
CD54	**Icam1**	**intercellular adhesion molecule 1**	**Yes**
CD55	**Cd55**	**CD55 molecule, decay accelerating factor for complement**	**Yes**
CD61	**Itgb3**	**integrin beta 3**	**Yes**
CD62e	Sele	selectin, endothelial cell	No
CD71	**Tfrc**	**transferrin receptor**	**Yes**
CD81	Cd81	CD81 antigen	No
CD93	Cd93	CD93 antigen	No
CD94	Klrd1	killer cell lectin-like receptor, subfamily D, member 1	No
CD98	Slc3a2	solute carrier family, member 2	No
CD102	Icam2	intercellular adhesion molecule 2	No
CD104	Itgb4	integrin beta 4	No
CD106	**Vcam1**	**vascular cell adhesion molecule 1**	**Yes**
CD117	**Kit**	**KIT proto-oncogene receptor tyrosine kinase**	**Yes**
CD119	**Ifngr1**	**interferon gamma receptor 1**	**Yes**
CD137	Tnfrsf9	tumor necrosis factor receptor superfamily, member 9	No
CD138	Sdc1	syndecan 1	No
CD144	Cdh5	cadherin 5	No
CD147	basigin	basigin	No
CD200	Cd200	CD200 antigen	No
CD284	Tlr4	toll-like receptor 4	No
CD309	Kdr	kinase insert domain protein receptor	No
Crry/p65	Cr1l	complement component (3b/4b) receptor 1-like	No
MadCam1	**MadCam1**	**mucosal vascular addressin cell adhesion molecule 1**	**Yes**
Meca32	Plvap	plasmalemma vesicle associated protein	No
PIR-A/B	**NA**	**NA**	**Yes**
Sca1	**Ly6a/e**	**lymphocyte antigen 6 complex, locus A/E**	**Yes**

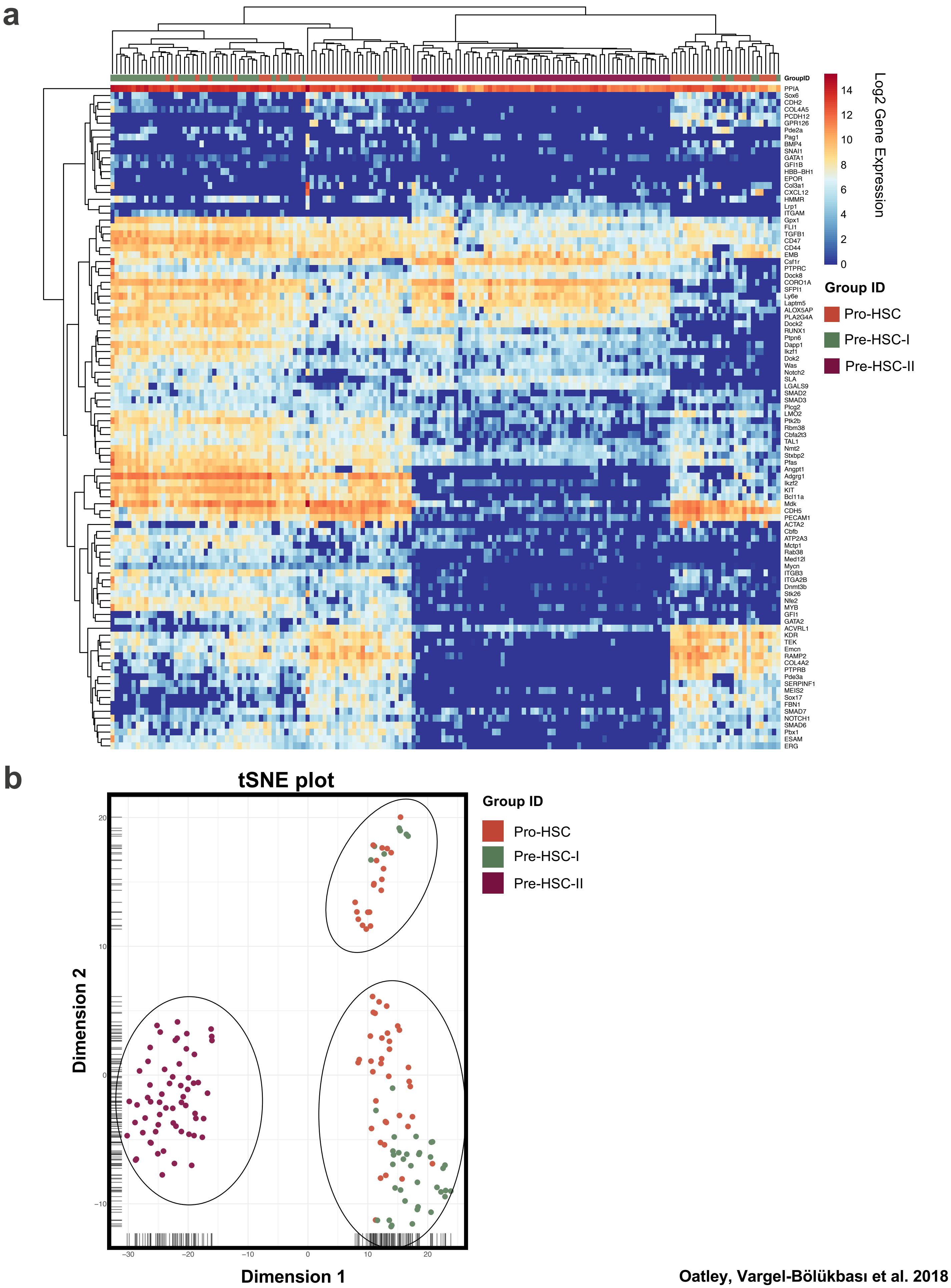
Supplementary Figure S2: Single-cell q-RT-PCR analysis of the four populations defined by CD44, VE-Cad and Kit



Supplementary Figure S2: Single-cell q-RT-PCR analysis of the four populations defined by CD44, VE-Cad and Kit

Single cells from indicated populations were isolated and tested for the expression of 95 genes by single-cell q-RT-PCR. The heatmap shows the result of the hierarchical clustering analysis (cells were clustered by Euclidian distance and the genes by Pearson correlation). Genes coding for *Gata2*, *Runx1*, *Lyl1*, *Lmo2*, *Tal1*, *Fli1* and *Erg* transcription factors are specifically co-expressed in the CD44^{Low}Kit^{Pos} population but not in the other two (see genes indicated by arrows). See also Figure 3 and Supplementary File S2.

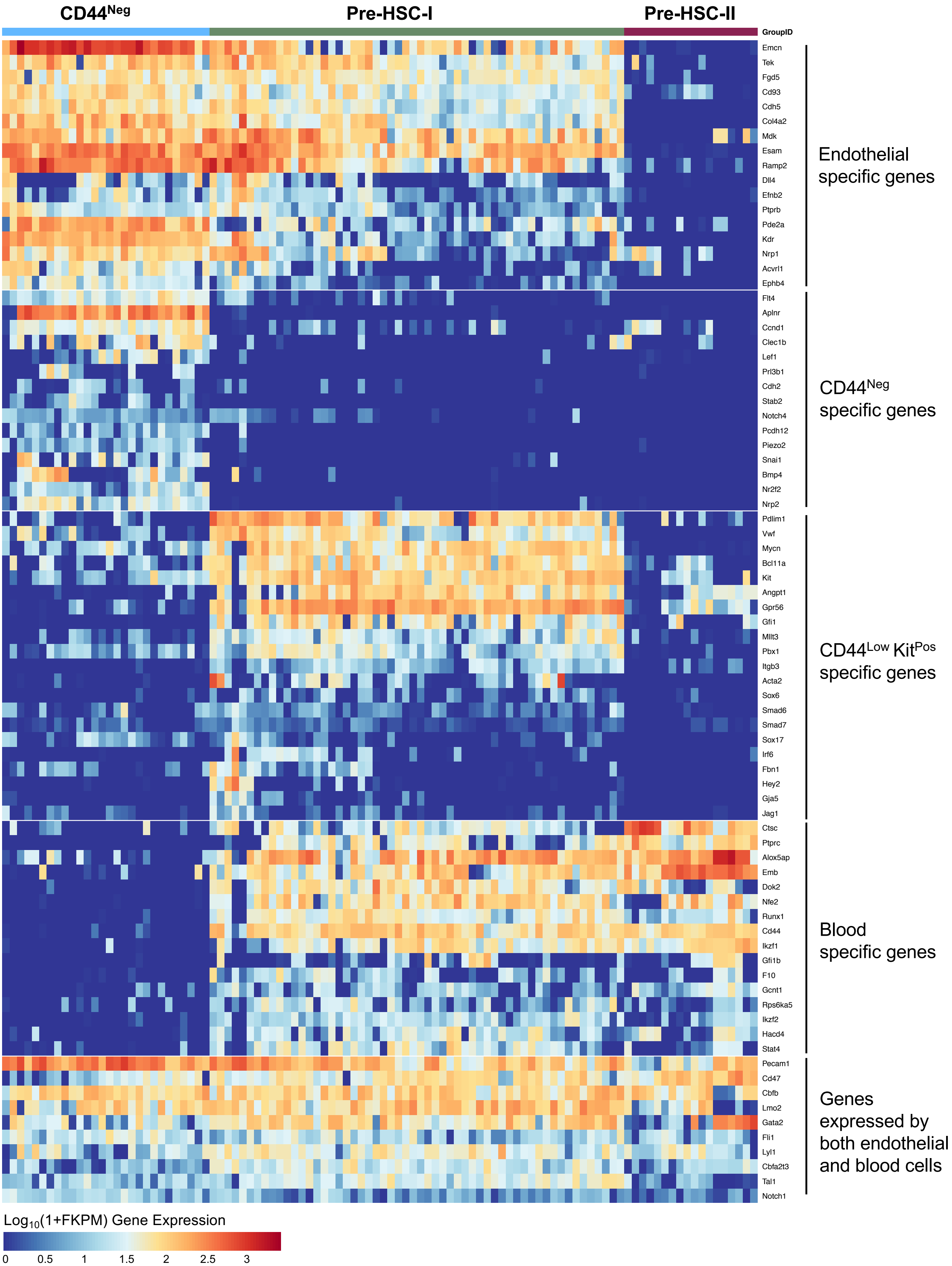
Supplementary Figure S3: Results of single cell q-RT-PCR analysis of Pro-HSC, Pre-HSC-I and Pre-HSC-II



Supplementary Figure S3: Results of single cell q-RT-PCR analysis of Pro-HSC, Pre-HSC-I and Pre-HSC-II

(**a**) Single cells from Pro-HSCs (VE-Cad$^+$ CD41$^+$ CD45$^-$CD43$^-$), Pre-HSC-I (VE-Cad$^+$ CD41$^+$ CD45$^-$ CD43$^+$), Pre-HSC-II (VE-Cad$^+$ CD45$^+$) populations were isolated and tested by single-cell q-RT-PCR. The heatmap shows the result of the hierarchical clustering analysis (cells were clustered by Euclidian distance and the genes by Pearson correlation). (**b**) tSNE plot from single cell q-RT-PCR data shown in (a). See also Figure 4 and Supplementary File S4.

Supplementary Figure S4: Analysis of the Zhou et al. sc-RNA-seq dataset

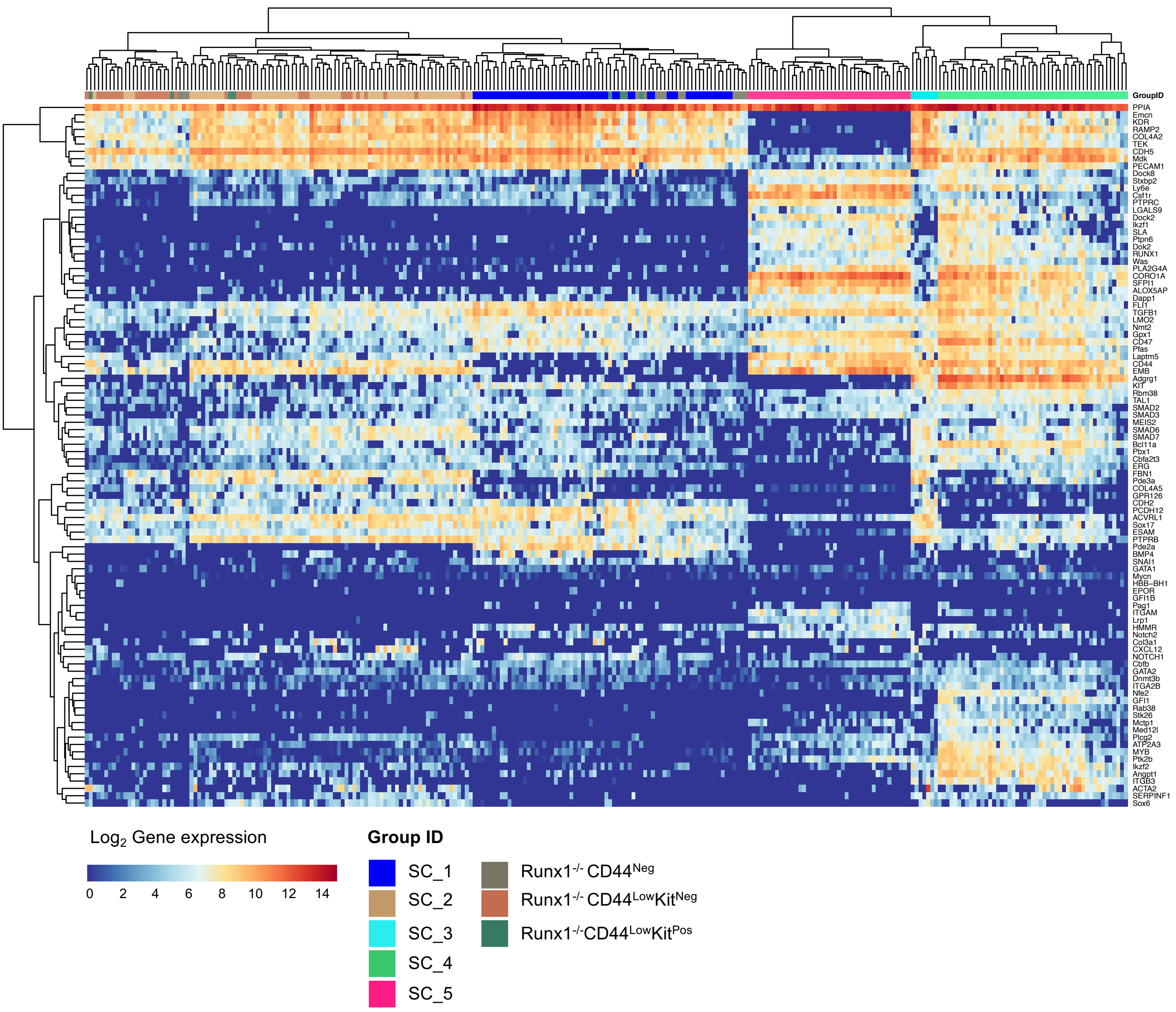


Oatley, Vargel-Bölükbası et al. 2018

Supplementary Figure S4: Analysis of the Zhou et al. sc-RNA-seq dataset

The heatmap shows the expression pattern of genes selected from Fig. 3 and Fig. 5 in the single cells studied by Zhou et al. [14] using sc-RNA-seq. These cells were isolated from E11 mouse embryos and included endothelial cells (CD44 Neg), Pre-HSC-I and Pre-HSC-II. The genes were grouped in the five indicated categories. See also Supplementary File S6.

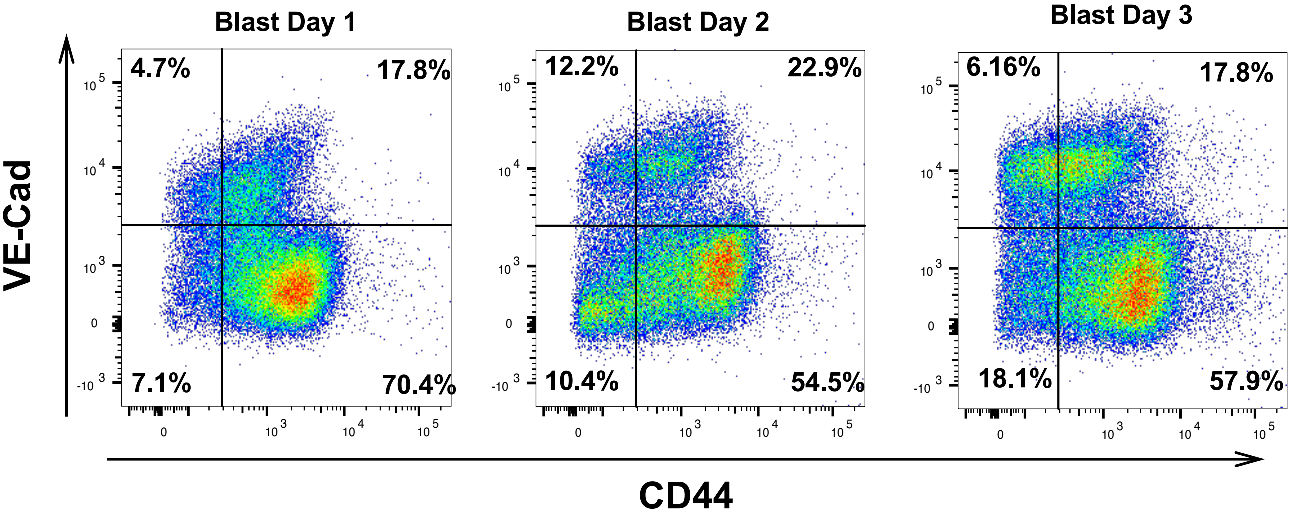
Supplementary Figure S5: Results of single cell q-RT-PCR analysis in the wild type and Runx1$^{-/-}$ AGM



Supplementary Figure S5: Results of single cell q-RT-PCR analysis in the wild type and Runx1$^{-/-}$ AGM

Single cells from Runx1$^{-/-}$ CD44^{Neg}, Runx1$^{-/-}$ CD44^{Low}Kit^{Neg} and Runx1$^{-/-}$ CD44^{Low}Kit^{Pos} populations were isolated and tested by single-cell q-RT-PCR. The heatmap shows the result of the hierarchical clustering analysis in combination with the wild type single-cells from Figure 3a (cells were clustered by Euclidian distance and the genes by Pearson correlation). See also Figure 7 and Supplementary File S7.

Supplementary Figure S6: Time course of CD44 expression during Haemangioblast culture



Supplementary Figure S6: Time course of CD44 expression during haemangioblast culture

Flow cytometry analysis of CD44 expression in Haemangioblast culture between day 1 and day 3. The dot plots show expression of VE-Cadherin and CD44 at the indicated time points. See also Figure 9.

Description of the supplementary files and tables

Supplementary File S1: Expression Data from single-cell RNA-seq from Fig. 1d

The first worksheet contains $\log_{10}$(1+TPM) expression data from single-cell RNA-seq and the second the metadata relative to the cells shown in Fig. 1d.

Supplementary File S2: Results of single-cell q-RT-PCR from Fig. S2

The first worksheet contains log2 expression data from single-cell q-RT-PCR and the second the metadata relative to the cells shown in Fig. S2.

Supplementary File S3: Results of single-cell q-RT-PCR from Fig. 3

The first worksheet contains log2 expression data from single-cell q-RT-PCR and the second the metadata relative to the cells shown in Fig. 3.

Supplementary File S4: Results of single-cell q-RT-PCR from Fig. S3

The first worksheet contains log2 expression data from single-cell q-RT-PCR and the second the metadata relative to the cells shown in Fig. S3.

Supplementary File S5: Results of the RNA sequencing from Fig. 5

First worksheet: Matrix showing rlog transformed expression values after normalization with the DSEQ2 package.

Second worksheet: Metadata related to the samples in Fig.5

Third worksheet: Gene list resulting from the differential expression analysis between the CD44^{Neg} and CD44^{Low}Kit^{Neg} populations (p-value_adjusted <0.01). The results were obtained following the Wald statistical test. Negative LogFC values indicate higher gene expression in CD44^{Low}Kit^{Neg} compared to CD44^{Neg} while positive LogFC values indicate higher expression in CD44^{Neg} compared to CD44^{Low}Kit^{Neg}.

Third worksheet: Expression matrix used in Fig. 5b.

Fourth worksheet: Expression matrix used in Fig. 5c.

Fifth worksheet: Expression matrix used in Fig. 5d.

Supplementary File S6: Expression Data from single-cell RNA-seq from Fig. S4

The first worksheet contains $\log_{10}$(1+TPM) expression data from single-cell RNA-seq and the second the metadata relative to the cells shown in Fig. 1d.

Supplementary File S7: Results of single-cell q-RT-PCR from Fig. S5

The first worksheet contains log2 expression data from single-cell q-RT-PCR and the second the metadata relative to the cells shown in Fig. S5.

Supplementary Table S1: Results of the antibody screen

List of the forty-two antigens (out of 176) expressed by VE-Cad^{+} cells from day 1.5 haemangioblast culture following the antibody screen. Sixteen of these markers have a bimodal expression (indicated in bold). See also Supplementary Figure S1.

Supplementary Table S2: List of primers for single-cell q-RT-PCR

These primers were used to detect the genes shown in Fig. 3, Fig. S3 and Fig. S5.

Supplementary Table S3: Results of the reporter metabolite analysis from Fig. 6

Table listing the results of the reporter metabolite analysis generated from the comparison of differentially expression genes between CD44^{Neg} and CD44^{Low}Kit^{Neg} populations.