ARTICLES
https://doi.org/10.1038/s41556-019-0445-8

nature cell biology

Multipotent *RAG1*+ progenitors emerge directly from haemogenic endothelium in human pluripotent stem cell-derived haematopoietic organoids

Ali Motazedian[1,2], Freya F. Bruveris[1,2], Santhosh V. Kumar[1,2], Jacqueline V. Schiesser[1,2], Tyrone Chen[3], Elizabeth S. Ng[1], Ann P. Chidgey[4], Christine A. Wells[3,5], Andrew G. Elefanty[1,2,4,6] and Edouard G. Stanley[1,2,4,6]*

Defining the ontogeny of the human adaptive immune system during embryogenesis has implications for understanding childhood diseases including leukaemias and autoimmune conditions. Using RAG1:GFP human pluripotent stem cell reporter lines, we examined human T-cell genesis from pluripotent-stem-cell-derived haematopoietic organoids. Under conditions favouring T-cell development, RAG1+ cells progressively upregulated a cohort of recognized T-cell-associated genes, arresting development at the CD4+CD8+ stage. Sort and re-culture experiments showed that early RAG1+ cells also possessed B-cell, myeloid and erythroid potential. Flow cytometry and single-cell-RNA-sequencing data showed that early RAG1+ cells co-expressed the endothelial/haematopoietic progenitor markers CD34, VECAD and CD90, whereas imaging studies identified RAG1+ cells within CD31+ endothelial structures that co-expressed SOX17+ or the endothelial marker CAV1. Collectively, these observations provide evidence for a wave of human T-cell development that originates directly from haemogenic endothelium via a RAG1+ intermediate with multilineage potential.

Studies of T-cell ontogeny suggest that zebrafish, mice and humans generate a wave of non-haematopoietic stem cell (HSC)-derived T-cell progenitors that precedes the establishment of the canonical thymus-dependent adult T-cell developmental pathway[1–6]. In the zebrafish, early T cells emerge directly from endothelium in the dorsal aorta[2], a domain encompassing a region functionally similar to the mammalian aorta-gonadnomesonephros (AGM) region (reviewed in[7,8]), which gives rise to HSCs in mice and humans[6,9]. Mouse studies suggest a variety of embryonic tissues that developmentally pre-date HSC formation can generate lymphocytes when exposed to lymphopoietic conditions[10–13]. Similarly, lymphoid-primed progenitors can be detected in the yolk sac, embryo proper, fetal liver and developing thymic rudiment before initiation of HSC expansion in fetal liver[4,14–17]. Collectively, these studies argue that the first lymphocytes are not derived from HSCs.

In the mouse, the thymus is first populated by a wave of progenitors marked by expression of *Rag1*, a gene that mediates T-cell receptor and B-cell immunoglobulin gene rearrangements[18,19]. These early Rag1+ progenitors, which possess T-cell and myeloid potential, are subsequently displaced by incoming HSC-derived progenitors[3]. Transcriptional profiling suggests that the earliest Rag1+ thymic immigrants are most similar to embryonic-day-11.5 fetal liver lymphoid progenitor cells[4]—a population that can be identified in fetal livers lacking haematopoietic repopulating activity.

Human haematopoietic ontogeny is thought to parallel that of the mouse (reviewed in[20,21]). Xeno-transplantation experiments show that, like the mouse[4], human fetal liver at Carnegie stage 16–17 contains unipotent T-cell progenitors before its colonization by HSCs[6]. The human fetal liver also harbours a population of CD19+ B-lineage cells at a similar developmental stage[5], providing support for the idea that human embryos execute an early, HSC-independent, lymphopoietic program. This conclusion is reinforced by human pluripotent stem cell (PSC) models that show human lymphoid lineages can be readily generated in vitro[22–26], despite the fact that robust production of HSCs is yet to be reported.

We previously described a haematopoietic differentiation system that generates AGM-like haematopoietic organoids from pluripotent stem cells[27]. These organoids robustly expressed NOTCH ligands on endothelial populations, thereby providing an environment permissive for the endothelial–haematopoietic transition that initiates definitive blood-cell development and for subsequent T-lineage differentiation[28–30]. In this study, we investigated the capacity of these organoids to generate lymphoid progenitors using RAG1:green fluorescent protein (GFP) reporter PSC lines. Through the combination of time-lapse imaging, flow cytometry, immunofluorescence analysis as well as single-cell-RNA-sequencing (scRNA-seq) analysis, we show that RAG1 marks multipotent T-cell progenitors that arise directly from haemogenic endothelium, potentially drawing together distinct observations pertaining to lymphocyte ontogeny spanning vertebrate species.

[1]Murdoch Children's Research Institute, Parkville, Victoria, Australia. [2]Department of Pediatrics, University of Melbourne, Parkville, Victoria, Australia. [3]Centre for Stem Cell Systems, Anatomy and Neuroscience, University of Melbourne, Parkville, Victoria, Australia. [4]Department of Anatomy and Developmental Biology, Monash University, Clayton, Victoria, Australia. [5]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. [6]These authors contributed equally: Andrew G. Elefanty, Edouard G. Stanley. *e-mail: ed.stanley@mcri.edu.au

Results

Generation of T-cell progenitors by PSC-derived organoids. To enable the identification of developing lymphoid cells, we developed and validated RAG1:GFP reporter PSCs, in which sequences encoding GFP were targeted to the RAG1 locus (Extended Data Fig. 1a–g). To generate haematopoietic progenitors, we adapted our spin embryoid body protocol[27] to a bulk cell-aggregation method[31], referred to here as swirler cultures. Embryoid bodies were subsequently transferred to air–liquid interface (ALI) cultures—a format previously shown to permit the maintenance of stromal elements, including endothelial cells, and to support the expansion of AGM-derived haematopoietic progenitors[9,32]. The relationships between the time line of differentiation, specific growth-factor treatments, culture formats and their morphological appearance are summarized in Fig. 1a–c. In this system, most cells expressed the pan-mesoderm marker CD13 (ref. [33]) by differentiation day 4 (Extended Data Fig. 2a); by day 8, 10–35% of the cells were CD34 bright and the majority of these were also DLL4+ (Extended Data Fig. 2a). At this stage, the cultures consisted of large cystic embryoid bodies (Fig. 1b, induced pluripotent stem cells (iPSCs); Extended Data Fig. 2b, human embryonic stem cells (hESCs)) and contained few CD43+ blood cells[27]. Transfer of day 8 embryoid bodies formed in swirler cultures to an ALI culture allowed the formation of organoids with an extensive network of vascular structures (Fig. 1c, iPSCs; Extended Data Fig. 2b, hESCs). By differentiation day 16, the cultures contained a substantial fraction of CD34+CD45+ blood-cell progenitors (Extended Data Fig. 2a).

The first RAG1+ cells could be visualized within CD31+ vascular structures from approximately differentiation day 16 (Fig. 1c, iPSCs; Extended Data Fig. 2b, hESCs). Early RAG1+ cells were CD43+CD45+CD41a− (Fig. 1d) and progressively acquired recognized markers of developing T cells (Fig. 1e). Cultures at later stages also contained double-positive (DP, CD4+CD8α+; CD8α is hereafter referred to as CD8) T cells, most of which were RAG1+. A substantial fraction of CD45+ blood cells were RAG1+ and expressed the pre-T-cell markers CD5 and CD7 at around differentiation day 32 (hESC, 35 ± 4%; iPSC, 29 ± 3%; Fig. 1f,g and Extended Data Fig. 2c). Using this system, CD56+CD7+ cells—indicative of natural killer (NK)-lineage differentiation—were only observed when the medium was supplemented with IL2 and IL15 from differentiation day 8 (Extended Data Fig. 2d). Overall, surface marker expression of the RAG1+ populations fitted well with the expected trajectory of T-cell development, with the initial downregulation of CD34 and progressive upregulation of CD5, CD7, CD4 and CD8 (Fig. 1h).

We also examined the potential of our organoid system to support B-lineage development by treating cultures with the NOTCH signalling inhibitor (γ-secretase inhibitor IX, (DAPT) from days 15 to 20. Given the requirement for NOTCH signalling to enable the endothelial-to-haematopoietic transition[22,34], DAPT was only added after the appearance of blood cells. Under these conditions—which also contained IL3, IL6 and G-CSF—a small number of RAG1+CD10+CD19+ pre-B cells was observed (Extended Data Fig. 2e), showing the potential of these organoids as a platform for probing B-cell development.

In the presence of IL7, T-lineage development was the predominant differentiation pathway supported by our organoid system. Although a proportion of cells reached the CD4+CD8+ stage, surface expression of CD3 was seldom observed, prompting us to question whether this block was intrinsic to the cells themselves or the organoid environment. To investigate this, we investigated whether organoid-derived RAG1+ cells had the capacity to generate CD3+TCR+ T cells under conventional T-cell differentiation conditions (OP9-DLL4; ref. [25]). To this end, RAG1+ cells were purified at differentiation day 28 and cultured on OP9-DLL4 stromal cells for 14 d. Flow cytometry confirmed the presence of CD3+TCRα/β+ cells that retained RAG1 expression (Fig. 2a,b). These results are consistent with the hypothesis that the limited supply of DLL4 provided by the organoid cultures is responsible for the failure of T-cell progenitors to express surface CD3.

Our organoid culture system also produced RAG1− cells that were CD5+CD7+ (hESCs, 65 ± 4%; iPSCs, 71 ± 3%), a fraction of which were single positive for either CD4 or CD8. To investigate this RAG1− population, we compared RAG1+ and RAG1− cells that were CD7+CD5+CD4−CD8− following two weeks of culture on OP9-DLL4 stromal cells. This analysis showed that a proportion of RAG1+ cells differentiated further to give CD4+CD8+CD3+ cells (Fig. 2c,d). Despite this, a significant fraction of cells lost expression of RAG1—around 85 ± 3% (n = 4) of cells were RAG1−CD5+CD7+ after two weeks. Conversely, RAG1− cells failed to generate CD4+CD8+ cells and no surface CD3 expression was detected in these cultures (Fig. 2d). These experiments suggest that developing T cells have a window to progress from the CD5+CD7+ stage to the CD4+CD8+ stage, with cells that miss this window becoming trapped at the CD5+CD7+ stage.

Organoid-derived T-cell progenitors have a typical T-cell developmental trajectory. Our RAG1:GFP reporter PSC lines provided an opportunity to correlate the expression of RAG1 with markers of T-cell development. Using the sorting strategy shown in Fig. 3a and the overall gating parameters shown in Extended Data Fig. 2f, we isolated specific RAG1+ populations, including double-negative (DN; RAG1+CD5+CD7+CD8−CD4−), immature single-positive (ISP; RAG1+CD8−CD4+) and DP stages, as well as cells that were RAG1−CD5+CD7+CD8−CD4− (RN). Analysis of these fractions confirmed that GFP expression increased as the cells progressed towards the DP stage (Extended Data Fig. 2g,h). Cells representing the CD3− and CD3+ fractions of DP+ cells from a single neonatal thymus were also examined to provide a frame of reference. These fractions were subject to RNA sequencing (RNAseq) analysis. The fidelity of the sorting strategy for PSC-derived samples is illustrated in Extended Data Fig. 2i.

Principle component analysis of data representing independent differentiation experiments indicated that like fractions clustered together (Fig. 3b). This RNAseq data also provided the opportunity to examine the relationship between the expression of RAG1 and GFP at the transcript level (Extended Data Fig. 2j). This analysis confirmed that the number of RAG1 and GFP reads were closely aligned and that the increase in expression of both genes paralleled the increase in GFP mean fluorescence intensity as cells progressed from CD5+CD7+ to CD4+CD8+ (Extended Data Fig. 2g,h,j).

A summary of the RNAseq analysis as it relates to T-cell development is provided in the form of a heat map (Fig. 3c). NOTCH signalling is critical for the early stages of T-cell differentiation[28–30] and our analysis shows that components of the NOTCH signalling pathway are upregulated at the ISP stage (Fig. 3c).

Human fetal thymocytes robustly express human leukocyte antigen (HLA) class II molecules[35], a property potentially linked to the involvement of these receptors in negative selection[36]. Consistent with this possibility, in our system, HLA class II was highest in cells in the DP stage (Fig. 3c). Chemokine signalling also plays an important role in thymocyte development[37]. Our data show modulation in chemokine-receptor expression as cells transit from the DN to the DP stage accompanied by upregulated expression of both *CCR9* and *CXCR4*. In the mouse, *Ccr9*, along with *Ccr7*, plays a role in thymic entry[38] and *Cxcr4* has been implicated in cortical localization and negative selection[39].

Examination of a cohort of archetypal T-cell-associated genes revealed two distinct patterns, one that includes genes whose expression peaks at the DN stage (for example, *RUNX3* and *GATA3*) and a larger set in which genes are progressively upregulated towards the DP stage (for example, *PTCRA* and *BCL11B*; Fig. 3c). A quantitative representation of the expression of a selected subset of these genes (shown in Fig. 3d) indicates that the trajectory

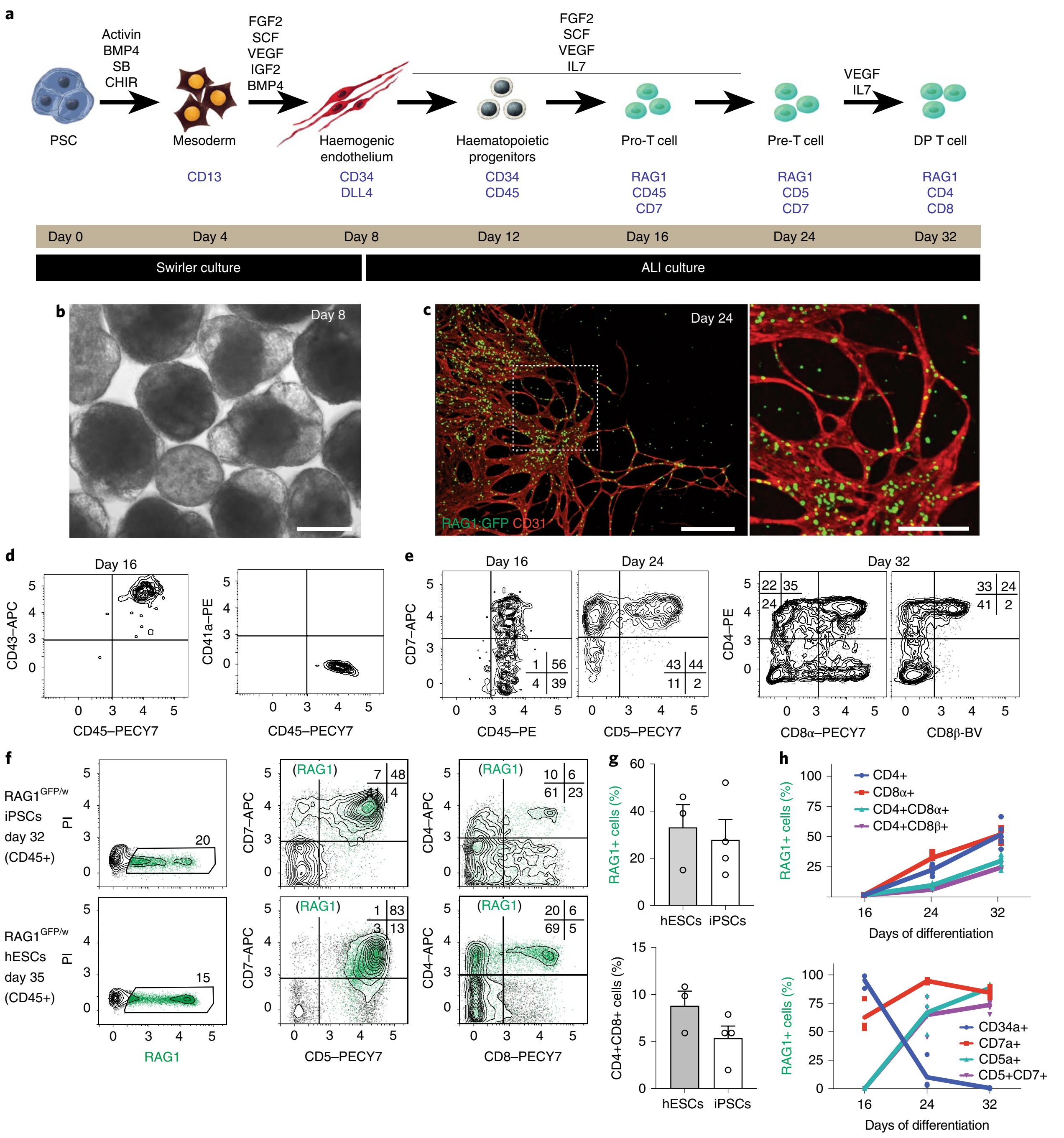


Fig. 1 | Generation and characterization of lymphopoietic organoids from PSCs. a, Schematic representation of the differentiation protocol including key identifiable immediate stages, growth-factor requirements and marker expression associated with each stage. **b**, Bright-field image of cystic embryoid bodies at differentiation day 8. **c**, Fluorescent image showing RAG1+ (GFP+) cells associated with vascular networks (CD31+) at differentiation day 24. The white box is magnified on the right. **d**, Flow cytometry analysis of RAG1+ cells at day 16 showing this population is uniformly CD43+CD45+CD41a–. **e**, Flow cytometry data from a single experiment documenting the evolution of a RAG1+ population over time, from the initial expression of RAG1 seen at differentiation day 16 to the upregulation of CD4, CD8α and CD8β observed at day 32. **f**, Flow cytometry analysis of cultures at differentiation day 32 (iPSCs; top) or 35 (hESCs; bottom) showing the relationship between and the relative proportion of CD45+ cells expressing RAG1, CD5, CD7, CD4 and CD8. **g**, Proportion of RAG1+ cells (top) in the cultures at day 32 and the fraction of CD45+ cells co-expressing CD4 and CD8 (bottom). The error bars represent the s.e.m. from $n=3$ (hESCs) and 4 (iPSCs) independent experiments. **h**, Graphical summary documenting the kinetics of T-cell marker expression on RAG1+ cells from day 16 to 32. The dots represent data from independent experiments using the iPSC-based RAG1$^{GFP/w}$ reporter line; the lines interconnect the mean value for the indicated markers at each time point. The number of individual data points for each marker and each day (d) are: CD34+, $n=3$ (d16), 4 (d24) and 2 (d32); CD7+, $n=3$ (d16), 4 (d24) and 6 (d32); CD5+, $n=2$ (d16), 4 (d24) and 6 (d32); CD5+CD7+, $n=1$ (d16), 4 (d24) and 6 (d32); CD4+, $n=1$ (d16), 4 (d24) and 5 (d32); CD8α+, $n=1$ (d16), 4 (d24) and 5 (d32); CD4+CD8β+, $n=1$ (d16), 3 (d24) and 3 (d32); CD4+CD8α+, $n=1$ (d16), 4 (d24) and 5 (d32). The numbers on the x and y axis of all flow cytometry plots are the exponents of the log$_{10}$[fluorescence] and the percentages of cells in each quadrant or associated with each gate are indicated; all of the cells displayed in green are RAG1+. The flow cytometry plots in **d**–**f** are representative examples from $n=4$ multiple experiments. Scale bars, 200 μm (**b** and **c**, right) and 400 μm (**c**, left).

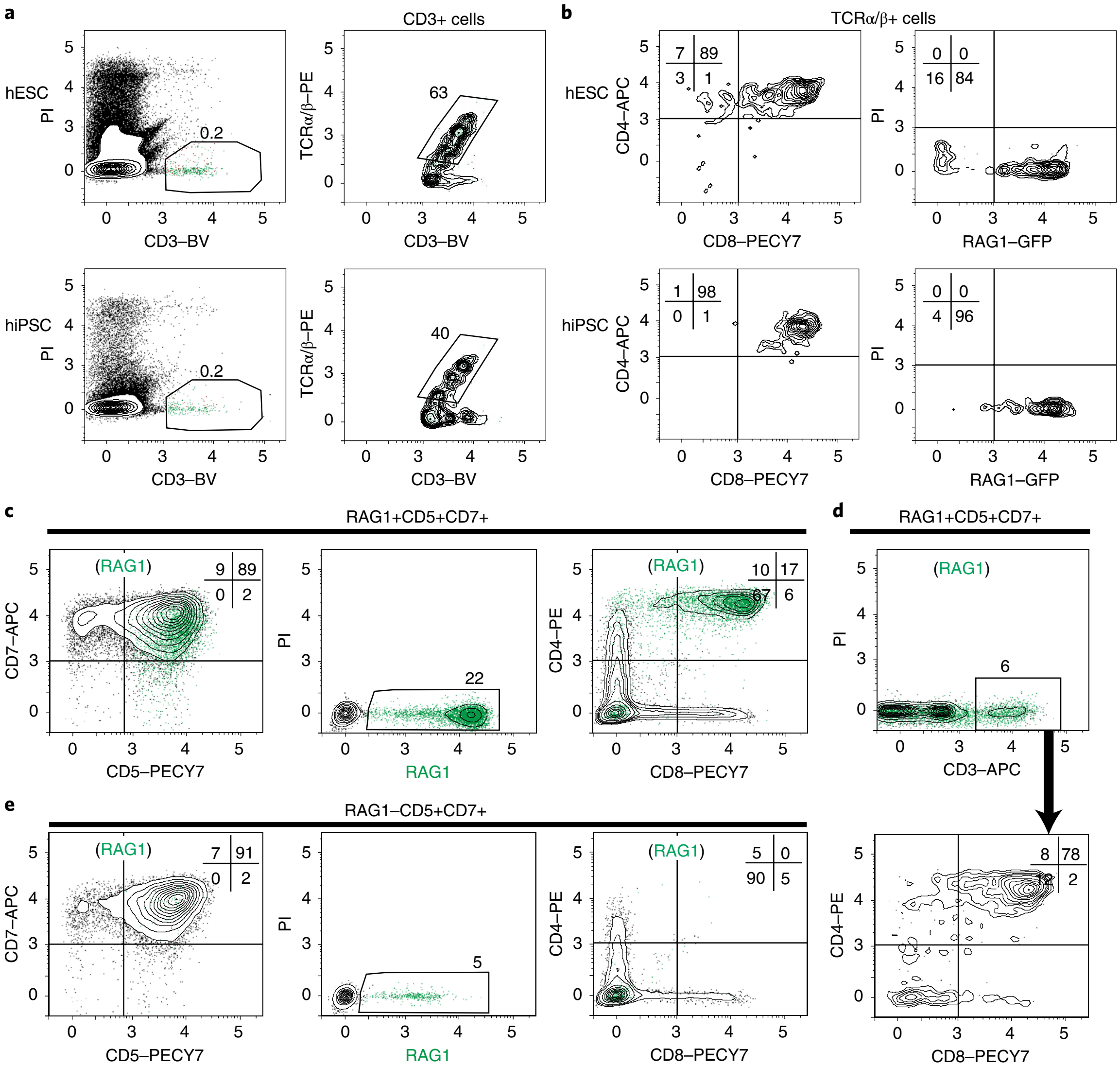


Fig. 2 | T-lineage-differentiation potential of RAG1+ cells. a, Flow cytometry analysis of organoid-derived RAG1+ cells sorted at differentiation day 28 and cultured on OP9-DLL4 stromal cells for 14 d post FACS purification. The flow cytometry plots show that approximately 50% of CD3+ cells also display TCRα/β on their surface at this stage. **b**, Flow cytometry plots indicating that the majority of CD3+ cells shown in **a** co-express CD4 and CD8 and retain robust expression of RAG1. **c**,**d**, Flow cytometry analysis showing that day 31 RAG1+CD5+CD7+ cells progress to CD4+CD8+CD3+ T cells when cultured for 14 d on OP9-DLL4 stromal layers. **e**, Flow cytometry analysis showing that a small proportion of day 31 RAG1−CD5+CD7+ cells can activate RAG1 expression but do not progress to the DP stage following co-culture with OP9-DLL4 stromal cells. The numbers on the *x* and *y* axes of all flow cytometry plots are the exponents of the $\log_{10}$[fluorescence] and the percentages of cells in each quadrant or associated with each gate are indicated. All of the cells displayed in green are RAG1+. The flow cytometry plots are representative examples drawn from multiple experiments; **a**,**b**, $n=4$ (hESCs) and 1 (iPSCs); **c**–**e**, $n=4$ (hESCs).

of differentiation for both the iPSC- and hESC-derived T-cell progenitors was similar (Fig. 3c,d).

Programmed cell death is a notable feature of T-cell development. Double-negative cells expressed high levels of *FAS* and *FASLG* along with *BCL2*, *CASP3* and *CASP4*, whereas DP cells expressed higher levels of genes encoding CASP7 and CASP8 along with *BCL2L1* (*BCL-XL*). The switch between these distinct combinations of apoptosis-related genes is consistent with the previously reported switch from IL7-dependent to TCR-dependent cell survival during T-cell ontogeny[40] (Fig. 3c).

Overall, this analysis indicates that organoid-derived RAG1+ T-cell progenitors undergo T-cell development characterized by the expression of typical stage-specific T-cell-associated gene sets. However, developmental progression was arrested at the DP stage, indicating the absence of appropriate environmental cues to allow the completion of positive and negative selection.

Early RAG1+ cells express markers associated with multiple haematopoietic lineages. To gain further insight into the developmental potential of the RAG1+ cells, we isolated early RAG1+ cells using fluorescence-activated cell sorting (FACS; see Fig. 4a and Extended Data Fig. 3a for examples of the overall gating strategy) and subjected a proportion of each population to functional assays and RNAseq analysis. Early RAG1+ populations isolated at day 20 contained a variable mixture of CD34+CD90+ (2.8 ± 2.4%) and CD34+CD90− (53 ± 7.2%) cells, with the balance of cells being CD34−CD90− (38 ± 6.6%; $n=4$; see Extended Data Fig. 3b for examples of experimental variability). As comparators, we analysed endothelial-enriched fractions (RAG1−CD34+CD90+; see Extended Data Fig. 4)—which also contained haemogenic endothelium and emerging blood progenitors—and haematopoietic (RAG1−CD34+CD90−) components from the same cultures (Fig. 4a).

As with the analysis of the T-lineage fractions (Extended Data Fig. 2j), we revisited the relationship between the expression of GFP and RAG1 by plotting the number of RNAseq reads for each gene in each fraction. This analysis showed that the GFP+ fractions were substantially enriched for reads representing RAG1 (and RAG2; Extended Data Fig. 3c). Plotting RNAseq reads mapping to RAG1 and GFP within the T-lineage and early RAG1 populations generated a curve (coefficient of multiple correlation (R^2) = 0.69), in which RAG1 (reads) = 0.822 GFP (reads) + 115 (Extended Data Fig. 3d). This curve suggested that GFP is likely to under-report RAG1 expression at low RAG1 expression levels and implied that a minimum level of RAG1 messenger RNA needed to accrue before the GFP transcripts were detectable.

For RNAseq analysis, samples were compared with cells representing the DN populations (CD5+CD7+CD4−CD8−RAG1+) presented in Fig. 3. Figure 4b shows a multidimensional-scaling plot indicating that, with the exception of the hESC-derived RAG1+ cells, samples representing specific sorting parameters clustered together. A summary of this analysis is presented in Fig. 4c; quantitative data representing selected genes are shown in Fig. 4d.

This analysis shows that the early RAG1+ cells expressed genes shared with the endothelial, myeloid and lymphoid populations (Fig. 4c). We had previously noted that expression of the *HOX* gene is rapidly downregulated as cells transit from endothelial to haematopoietic cells[27]. This relationship held true in our current experiments, except for the persistent expression of *HOXB1–4* in the most-differentiated CD5+CD7+RAG1+ population. In addition, the average expression of the 5′ HOXA cluster genes *HOXA9–13* in the early RAG1+ cells aligned most closely with that of cells in the CD34+CD90+RAG1− fraction, thereby suggesting a close developmental juxtaposition of these two populations (Extended Data Fig. 3e). This relationship was also reinforced by the residual expression of CD34 and CD90 in early RAG1+ cells (Fig. 4d). Similarly, although there is a large cohort of endothelial-associated genes, whose expression is markedly reduced in all of the blood related fractions (*NOTCH4* through to *CDH5* (*VECAD*) in Fig. 4c), other genes such as *NOTCH1*, *NOTCH3*, *CD34* and *THY1* (*CD90*) show persistent expression in many of the individual early RAG1+ samples. The co-expression of other genes (*JAG1* through to *GATA2* in Fig. 4c) highlights the similarities between the endothelial-enriched RAG1−CD34+CD90+ and early haematopoietic cells (RAG1−CD34+CD90−), and distinguishes them from the early RAG1+ cells and RAG1+CD5+CD7+ populations.

The RAG1−CD34+CD90+, RAG1+ and RAG1−CD34+CD90− populations also share the expression of erythro-myeloid lineage-associated genes (*NFE2* through to *GFI1B* in Fig. 4c). The elevated expression of these genes in the RAG1+ fraction is consistent with the possibility that this population may harbour a subset of cells with erythro-myeloid potential, a property previously identified in the early Rag1+ cells from the mouse fetal liver[41].

In addition to progenitor-related genes, there was a larger group of erythro-myeloid lineage-associated genes (*LIF* through to *CITED2*, including *GATA1* and *CD14*, Fig. 4c) that was upregulated in the RAG1−CD34+CD90− fraction, and expression of some of these was also elevated in the RAG1−CD34+CD90+ samples representing endothelium/stem progenitor cells. Genes within this group were downregulated in the RAG1+ fraction, thereby indicating a potential inverse correlation between the expression of RAG1 and erythro-myeloid commitment. As noted above, such a correlation has been previously observed in studies of Rag1+ cells in the mouse fetal liver[17,41].

The expression of many genes associated with T-cell development was elevated in the RAG1+ fraction and the clearly T-cell-committed CD5+CD7+RAG1+ fractions. Although the expression of most of these increased with further differentiation (*CD3E* through to *PTCRA* in Fig. 4c), a small number seemed to be highly expressed in the early RAG1+ fraction, including *RAG1* itself as well as *RAG2* and *CD1E*.

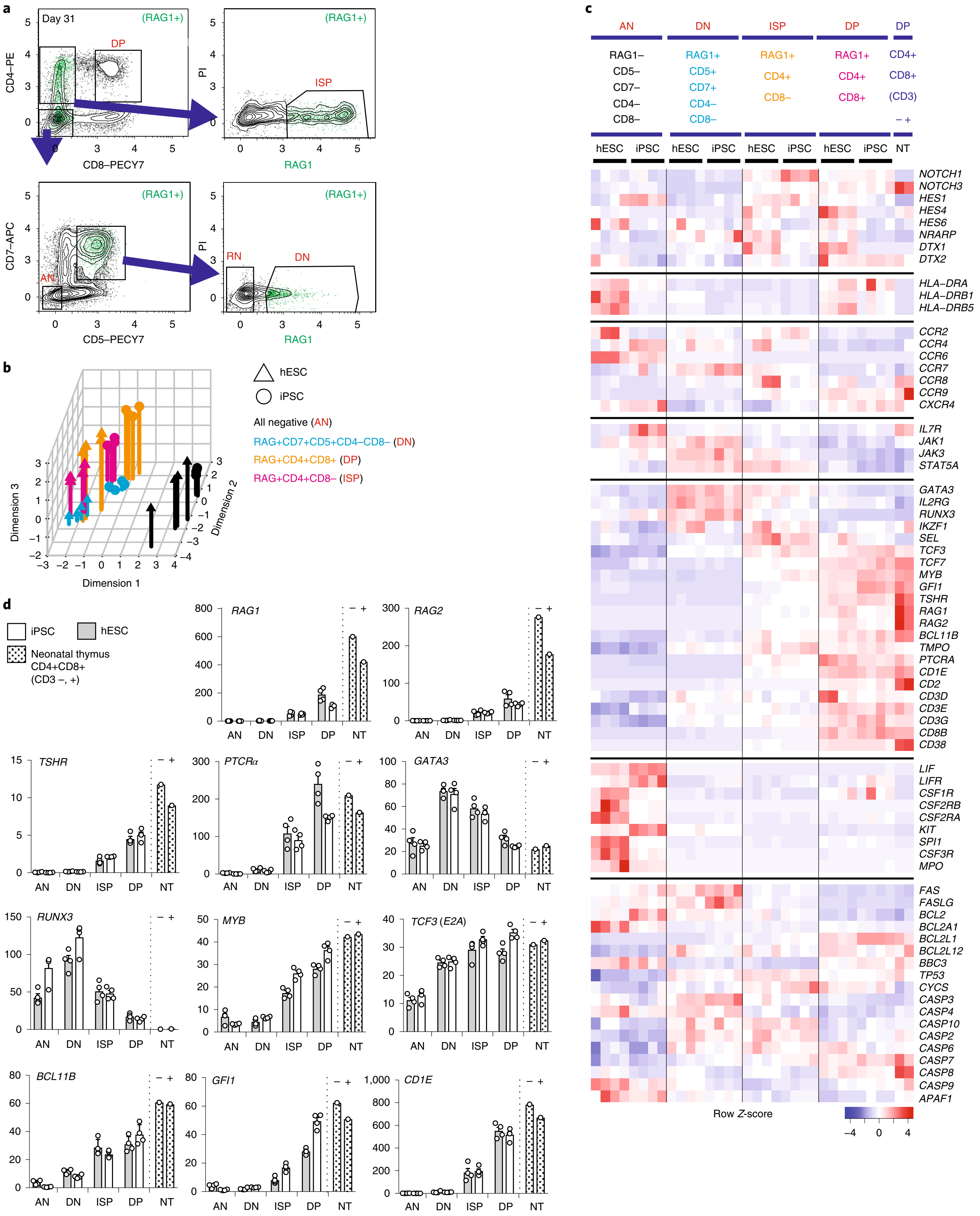
To provide a second independent measure of the differentiation stages captured in the early RAG1+ population, we also compared the expression of T-cell-receptor genes in the RAG1−CD34+CD90+, early RAG1+ (as mentioned in Fig. 4a), and DN, ISP and DP populations (as mentioned in Fig. 3a). This analysis showed that the iPSC and hESC lines had similar differentiation trajectories (Extended Data Fig. 4a,b). Overall, this set of experiments indicated that, as a population, early RAG1+ cells shared characteristics of both the endothelial and stem progenitors as well as erythro-myeloid and T-cell progenitors.

Early RAG1+ cells have myeloid, erythroid and lymphoid potential. We examined the differentiation potential of the early RAG1+, RAG1−CD34+CD90 and RAG1−CD34+CD90+ cells isolated for the RNAseq experiments described in Fig. 4. Sorted cells were plated into methylcellulose to assess the colony-forming potential or seeded onto OP9 or OP9-DLL4 stromal cells to assess B-cell or T-cell capacity. These experiments indicated that the RAG1+ pool was highly clonogenic and generated colonies that had the morphological appearance of dispersed myeloid and compact erythroid colonies (Fig. 5a,b, iPSC; Extended Data Fig. 5a, hESC). Flow cytometry analysis of methylcellulose cultures at day 12 showed that cells derived from all three sorted fractions expressed CD235a (glycophorin A), which is indicative of erythroid lineage cells (Fig. 5c,d). Similarly, 10–40% of cells expressed the myeloid marker CD14, with a proportion of these also expressing the macrophage marker CD16 (Fig. 5c). Quantitation of five experimental replicates derived from the differentiation of RAG1$^{GFP/w}$ iPSCs showed that all fractions had a propensity to generate myeloid lineages marked by CD14, whereas the RAG1−CD34+CD90− cells were biased towards generating erythroid lineages (Fig. 5d).

The myeloid differentiation capacity of the RAG1+ cells was further confirmed by analysis of OP9 stromal co-cultures. These experiments

Fig. 3 | Organoid-derived T-cell progenitors traverse established developmental stages. a, Flow cytometry sorting strategy used to isolate RAG1+ T-cell progenitors expressing combinations of CD5, CD7, CD4 and CD8 at differentiation days 31–32. The specific parameters used for each fraction are listed in **b** and across the head of panel. The numbers on the *x* and *y* axes are the exponents of the $\log_{10}$[fluorescence]. All of the cells displayed in green are RAG1+. The flow cytometry plots are representative examples drawn from *n* = 4 independent experiments. **b**, Multidimensional-scaling plot showing the relationship between the RNAseq data representing the indicated fractions derived from *n* = 4 independent differentiation experiments conducted using the hESC and iPSC RAG1$^{GFP/w}$ reporter cell lines. **c**, Heat map representation of RNAseq gene expression analysis of the fractions indicated in **b** as well as CD4+CD8+ CD3+ and CD3− human neonatal thymocytes (NT) derived from a single donor. The expression levels were normalized to the average expression of each gene across all samples. The row *Z*-score indicates the fold change in gene expression relative to the row average. **d**, RNAseq gene expression profiles of selected genes across the indicated fractions. The *y* axis shows the data expressed as reads per kilobase per million (RPKM). The bar height shows the mean for each sample; the open circles represent the individual data points that contributed to the calculation of the mean. With the exception of the human neonatal thymus sample (*n* = 1), all graphs represent data from *n* = 4 independent differentiation experiments.

indicated that both RAG1−CD34+CD90+ cells and RAG1+ fractions efficiently gave rise to myeloid precursors (CD16+; Fig. 5e), and that the RAG1−CD34+CD90+ and, to a lesser extent, RAG1+ populations gave rise to RAG1+CD10+CD19+ B-cell progenitors (Fig. 5e). A similar spectrum of potential was also observed for hESC-derived RAG1+ cells (Extended Data Fig. 5b,c). As expected,



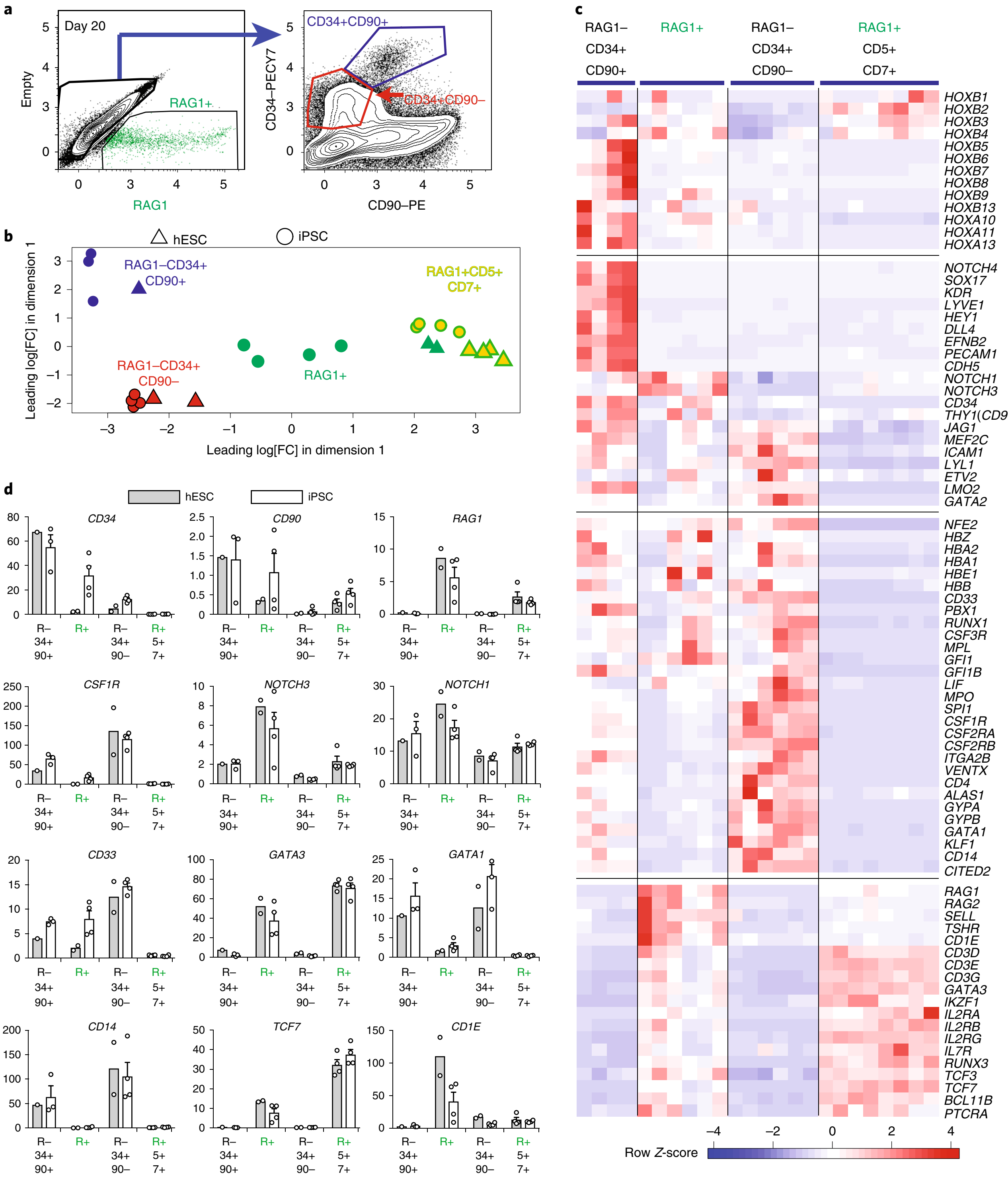


Fig. 4 | Early RAG1+ progenitors express genes associated with endothelial, erythroid and myeloid lineages. a, Flow cytometry sorting strategy for the isolation of the indicated populations from differentiating cultures of RAG1$^{GFP/w}$ iPSCs at differentiation days 20–21. The numbers on the *x* and *y* axes are the exponents of log$_{10}$[fluorescence]. All of the cells displayed in green are RAG1+. The flow cytometry plots are representative examples drawn from multiple experiments; $n = 4$ (iPSCs) and 2 (hESCs). **b**, Multidimensional-scaling plot showing the clustering of RNAseq data generated from the indicated cell populations from two independent RAG1$^{GFP/w}$ PSC reporter lines. Four independent experiments were performed with iPSCs and two with hESCs. The CD5+CD7+RAG1+ data are derived from Fig. 2 and represent separate differentiation and cell sorting experiments. FC, fold change. **c**, Heat map representation of gene expression data across the indicated fractions showing that the RAG1+ population expresses a number of genes indicative the endothelial, erythroid and myeloid lineages. **d**, Quantitative representation of the expression levels of the indicated genes derived from RNAseq data representing the iPSC and hESC RAG1$^{GFP/w}$ reporter lines. The *y* axis shows the data expressed as RPKM. The number of samples shown for each fraction are: RAG1–CD34+CD90+ (R–34+90+), $n = 3$ (iPSC) and 1 (hESC); RAG1+ (R+), $n = 4$ (iPSC) and 2 (hESC); RAG1–CD34+CD90– (R–34+90–), $n = 4$ (iPSC) and 2 (hESC); and RAG1+CD5+CD7+ (R+5+7+), $n = 4$ (iPSC and hESC). The error bars represent the s.e.m.; the bar height shows the mean for each sample.

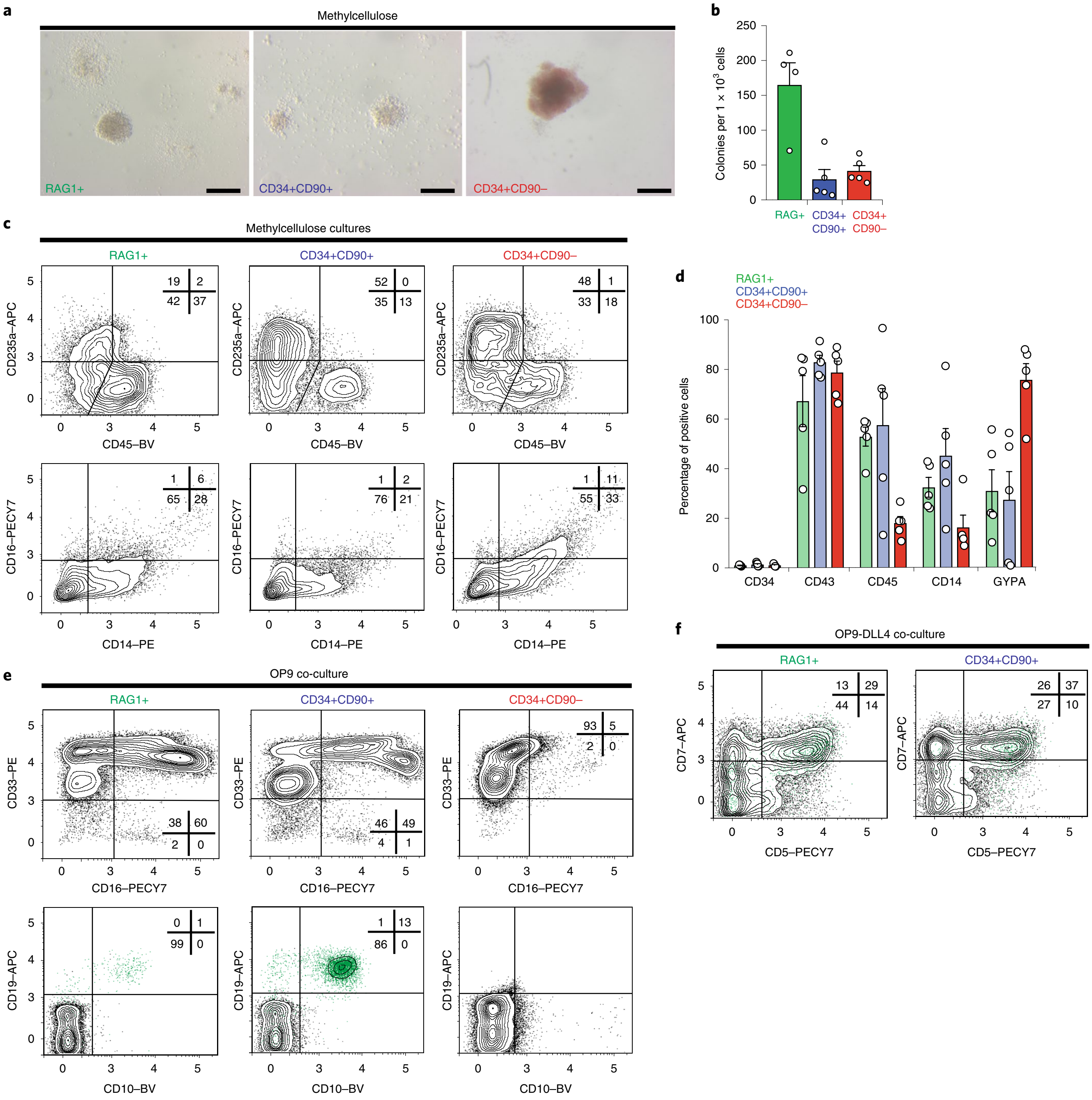


Fig. 5 | PSC organoid-derived RAG1+ cells are multipotent. a, Bright-field low-magnification images of blood-cell colonies formed after 10 d in methylcellulose cultures seeded with flow-sorted RAG1+ and RAG1− (CD34+CD90+ and CD34+CD90−) cells. Scale bars, 400 µm. Examples of diffuse myeloid colonies and variably haemoglobinized compact erythroid colonies are shown. **b**, Total number of methylcellulose colonies derived from the three indicated fractions. The mean and s.e.m. are shown; $n=4$ (RAG1+) and 5 (CD34+CD90+ and CD34+CD90−) independent experiments. **c**, Flow cytometry analysis of cells harvested from methylcellulose cultures at day 12, showing the presence of erythroid (CD235a/glycophorin A) and myeloid (CD14 and CD16) cells. **d**, Summary of the flow cytometry analysis of cells derived from methylcellulose cultures seeded with the indicated fractions. The mean and s.e.m. are shown; $n=5$ independent experiments for each marker. **e**,**f**, Flow cytometry analysis of CD45+ cells derived from the indicated fractions cultured on OP9 stromal cells (**e**) for 4 weeks and OP9-DLL4 stromal cells (**f**) for 2 weeks, showing that RAG1+ cells give rise to CD14+ myeloid, CD5+CD7+ T-cell progenitors and, to a lesser extent, CD10+CD19+ B-lymphoid-lineage cells. The numbers on the *x* and *y* axes of all flow cytometry plots are the exponents of the $\log_{10}$[fluorescence] and the percentages of cells in each quadrant are indicated; all of the cells displayed in green are RAG1+. The flow cytometry plots are representative examples drawn from multiple experiments; $n=5$ (**c**), 4 (**e**) and 2 (**f**).

both the RAG1−CD34+CD90+ and RAG1+CD34+CD90− fractions efficiently generated CD5+CD7+ T-cell progenitors in OP9-DLL4 co-cultures (Fig. 5f). Collectively, this set of experiments indicates that RAG1 marks early haematopoietic progenitors with differentiation potential encompassing the erythro-myeloid and lymphoid lineages.

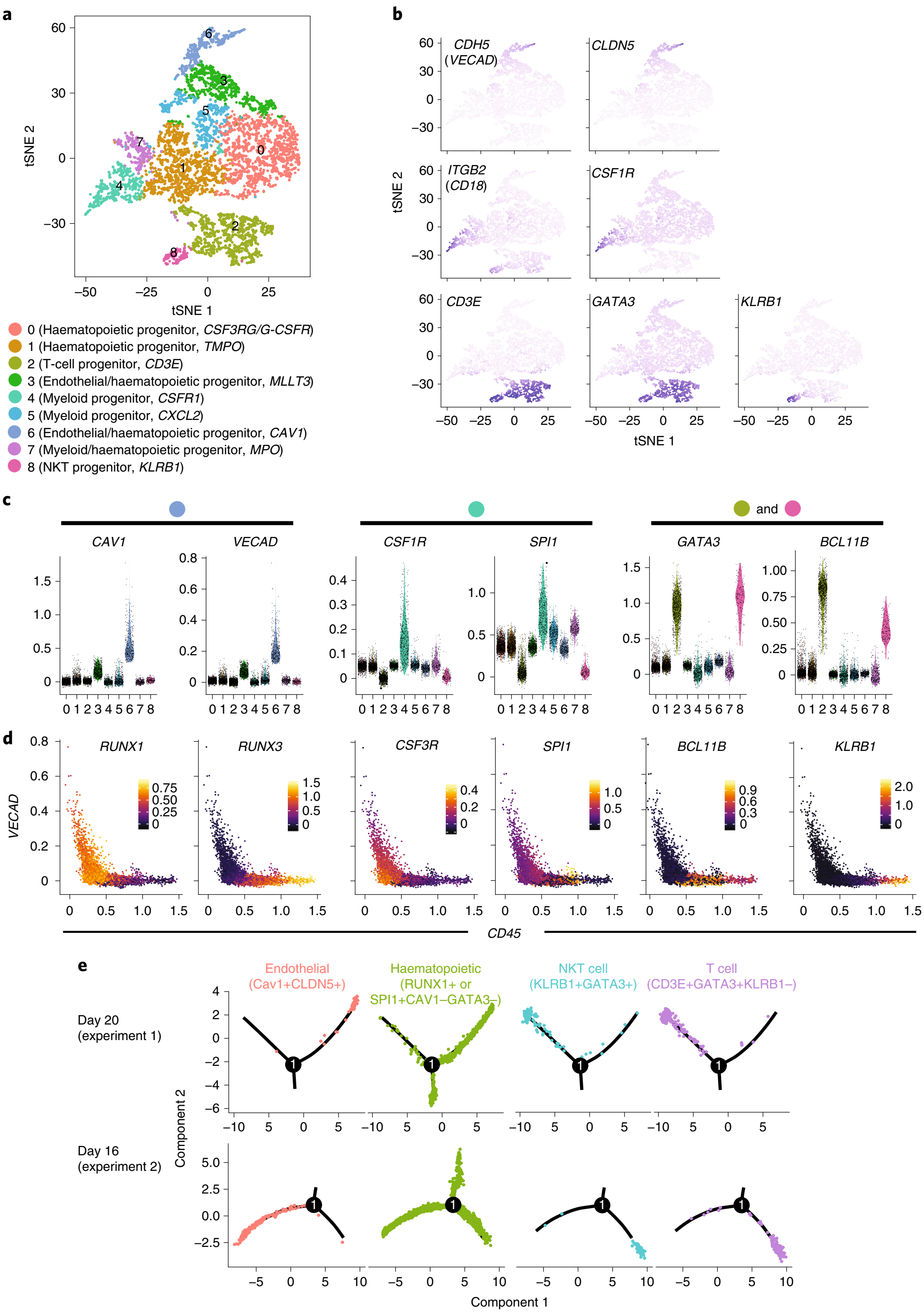


Fig. 6 | Early RAG1+ populations from cultures at differentiation day 20 contain discrete cell subsets enriched for endothelial, myeloid and T-cell markers. a, Relative positions of the indicated populations across two dimensions in a tSNE plot of scRNA-seq data derived from 6,347 cells. The naming of each population was based on the expression of a key gene(s) in that subset, as indicated. **b**, Relative expression levels of the indicated genes across the whole early RAG1+ population shown in tSNE plots. The cells expressing the highest level of each gene have the darkest colour. **c**, Distribution of expression levels in individual cells for the indicated genes in each population. The *x* axis shows the population number and the *y* axis shows the level of expression (ln[TPM + 1]; TPM, transcripts per million). Each point represents an individual data point (that is, $n > 10$). Cell populations are numbered and colour coded as per the legend in **a**. **d**, Dot plot representation of the expression levels of the indicated genes in cells expressing VECAD (nominally endothelium) and/or CD45 (nominally haematopoietic). The relative expression for the indicated gene is shown by colour coding each cell from highest (yellow) to lowest (purple) as a function of the ln[TPM + 1] value of that gene for each cell type. Values on *x* and *y* axes show the level of expression (ln[TPM + 1]) for the indicated gene. **e**, Pseudo-time analysis of cells defined by the indicated gene expression signatures using the Monocole algorithm[38] showing a single branch point from a single differentiation trajectory. Each dot represents a single cell and each branch indicates a potential cell fate defined by the gene signature as indicated.

RAG1+ cells emerge directly from the endothelium of haemogenic organoids. To explore the heterogeneity of the early RAG1+ population, we performed two single-cell sequencing experiments on RAG1+ cells isolated from day 20 (Fig. 6, Extended Data Figs. 6 and 7a) and 16 (Extended Data Fig. 7b–f) cultures of RAG1$^{GFP/w}$ iPSCs, using the sorting parameters shown in Fig. 4a. Analysis of this data identified several subpopulations that could be related to the bulk sequencing data described earlier (Fig. 4) and the assays for multipotentiality (Fig. 5). These populations were evident in the raw data (Extended Data Fig. 7a,b for sorted cells from both day 20 and 16) and more visible following diffusion-based imputation using Markov affinity-based graph imputation of cells (MAGIC[42]; Fig. 6a–d, sorted day 20 cells). A nominal identity for each population was designated based on gene sets identified using the MAGIC-transformed data, which are summarized as a heat map in Extended Data Fig. 6. These groups are also reflected in the untransformed data from the sorted day 16 cells (Extended Data Fig. 7c–f).

This analysis suggests that the early RAG1+ population included cell types that had characteristics of endothelial cells, and myeloid- and T-lineage cells, the latter of which also include a small sub-group with NKT-like characteristics. T-distributed stochastic neighbour embedding (tSNE) plots for key genes in this progression are shown in Fig. 6a,b, with a quantitative representation of selected genes presented in the accompanying violin plots (Fig. 6c). Our interpretation of this data is that early RAG1+ cells represent a transient population, which originates from an endothelial precursor and includes cells with the characteristics of progenitors representing multiple lineages. The validity of this interpretation is supported by plots displaying the expression of specific genes in relation to that of *VECAD* (*CDH5*; endothelial/haematopoietic stem progenitor) and *CD45* (haematopoietic; Fig. 6d). In this transition, intermediate populations expressing the myeloid markers *CSF3R* (*G-CSFR*/*CD114*) and *SPI1* (*PU.1*) are also apparent, and the NKT marker *KLRB1* in cells with the highest levels of *CD45* transcripts. The overall conclusions from this data are supported by a second independent single-cell sequencing experiment that examined RAG1+ cells isolated at differentiation day 16 (Extended Data Fig. 7b).

Based on marker genes representative of specific cell clusters derived from the above analyses, four groups of cells including endothelial (CAV1+CLDN5+), haematopoietic (SPI1+ or RUNX1+GATA3−CAV1−), T-lineage (GATA3+CD3E+KLRB1−) and NKT-lineage (GATA3+KLRB1+) progenitors were selected for single-cell trajectory analysis using the Monocole algorithm[43] (Fig. 6e). This analysis shows that the haematopoietic progenitors branch off midway, possibly indicative of cells acquiring a non-T-lymphoid fate.

A similar pseudo-time relationship was observed for cells from experiment 2, isolated at differentiation day 16. Moreover, the haematopoietic branch observed in both experiments suggests that some RAG1+ cells are primed for a non-T-lymphoid fate during the early stages of differentiation and that RAG1 marks a heterogeneous population at these early time points.

Overall, the scRNA-seq data are consistent with our gene expression analysis of the early RAG1+ populations and the cell potentiality experiments, which provides support for the concept that early RAG1+ populations contain cells with erythroid and myeloid potential, mirroring the characteristics of similar cells found in the mouse fetal liver[17,43].

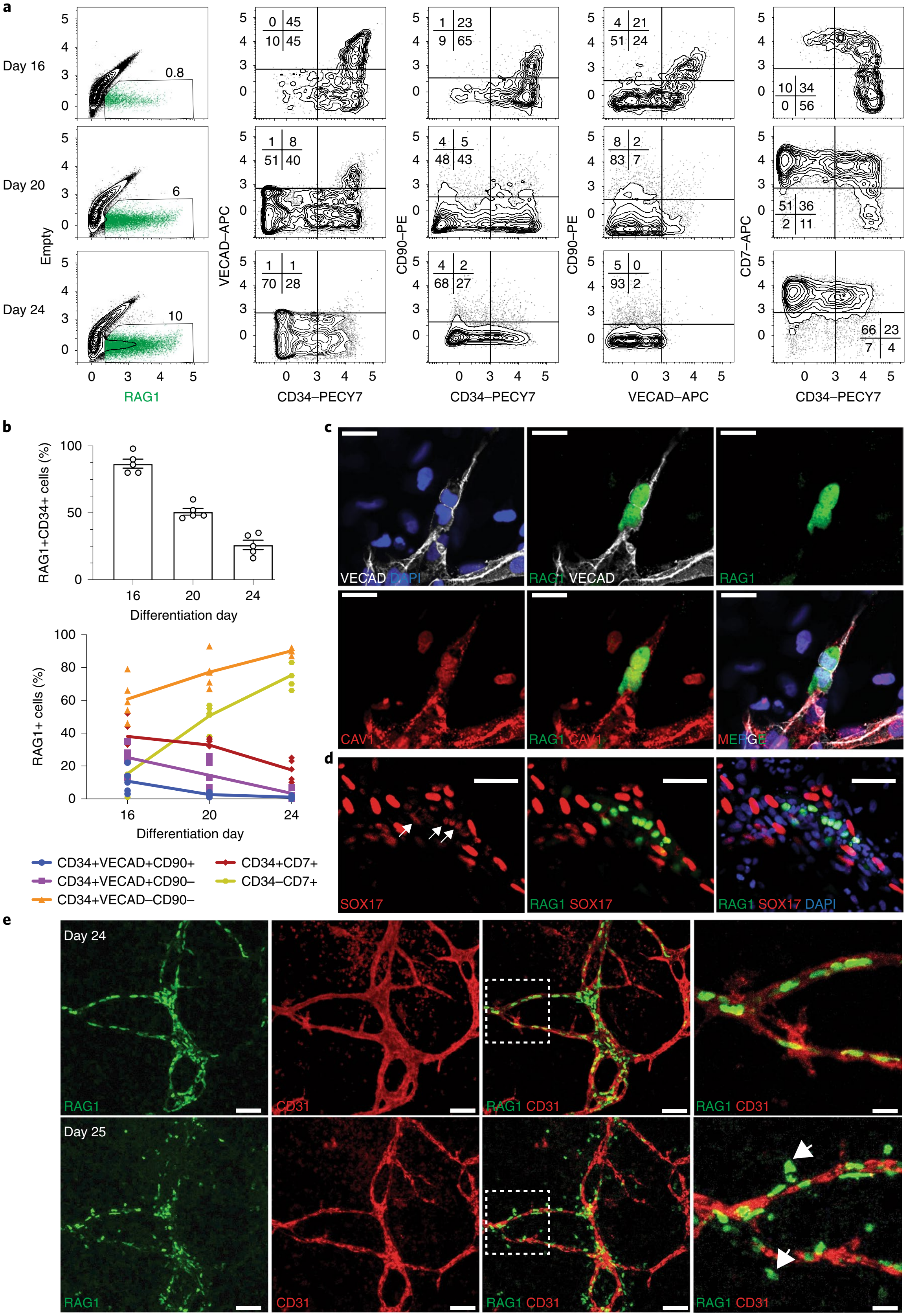
We were particularly interested in the possibility that RAG1+ cells emerged directly from haemogenic endothelium. To examine this possibility, we performed flow cytometry and immunofluorescence analysis of differentiation cultures that had just initiated RAG1 expression and also re-assessed the differentiation profile of RAG1+ cells that co-expressed the progenitor markers CD34, VECAD and CD90. Consistent with the RNAseq data, flow cytometry analysis showed that a subset of RAG1+ cells at differentiation day 16 expressed the endothelial/haematopoietic progenitor markers CD34, VECAD (CDH5) and CD90 (Fig. 7a, iPSC; Extended Data Fig. 8a,b, hESC). Around half the RAG1+ cells were negative for CD7 (Fig. 7a,b, iPSC; Extended Data Fig. 8c, hESC), thereby confirming the results of earlier experiments (Fig. 1e,h). The proportion of RAG1+ cells expressing progenitor markers decreased over the ensuing days, with RAG1+ cells progressively upregulating CD7 by differentiation day 24 (Fig. 7a,b).

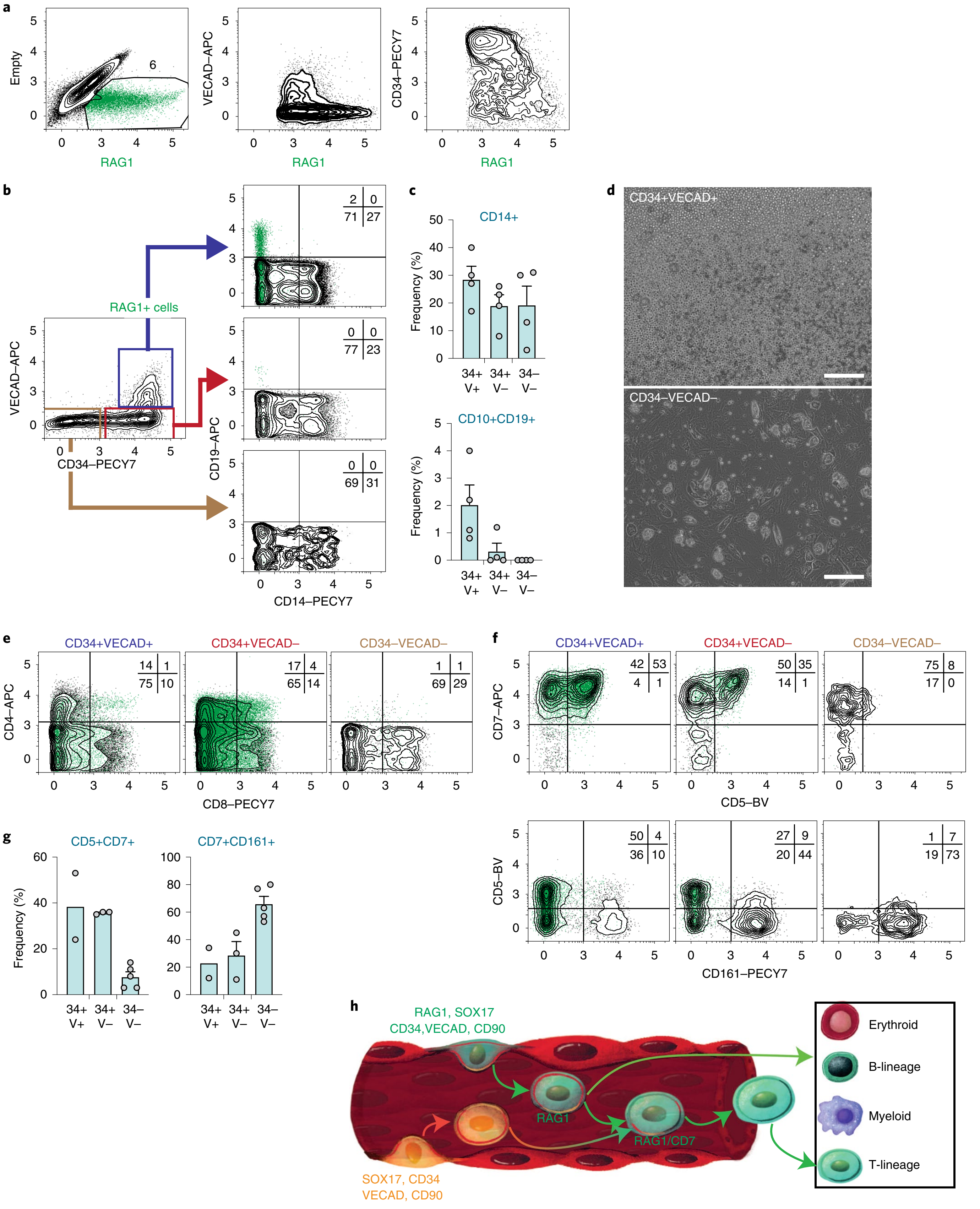
The close relationship between RAG1 expression and endothelial cells was reinforced by immunofluorescence analysis, which showed that RAG1+ cells co-expressed VECAD and the endothelial marker caveolin-1 (CAV1; Fig. 7c, Extended Data Fig. 8d and *Z*-stack Supplementary Video 1). On rare occasions, we also observed the co-expression of RAG1 with the haemogenic endothelial marker SOX17 (Fig. 7d and *Z*-stack Supplementary Video 3), a gene whose expression is rapidly lost as cells undergo the endothelial–haematopoietic transition[27]. As such, RAG1+ cells were initially closely aligned with CD31+ vascular structures (Fig. 7e), which also served as ‘highways’ along which the RAG1+ cells travelled (Supplementary Videos 4–6). This migratory phenotype was a defining characteristic of T-cell progenitors generated in haematopoietic organoids.

Although the majority of early RAG1+ cells were uniformly CD34+, the levels of VECAD expression showed an inverse relationship with those of RAG1 (Fig. 8a). We revisited the relationship between the expression of RAG1, progenitor markers and potentiality, examining the capability of RAG1+CD34+VECAD+, RAG1+CD34+VECAD− and RAG1+CD34−VECAD− cells to generate myeloid-, B- and T-lineage cells in stromal co-cultures to detail this differentiation trajectory (Fig. 8b–g). Although all populations produced CD14+ myeloid cells with similar efficiencies, B-lineage potential was almost entirely restricted to the CD34+VECAD+ fraction (Fig. 8c). In addition, we observed that the CD34+VECAD+ fraction harboured cells with the highest proliferative potential, which was evident by the overcrowding of cultures during the two-week co-culture period (Fig. 8d).

Fig. 7 | RAG1+ cells emerge directly from haemogenic endothelium. a, Time-course analysis of the expression of CD34, VECAD, CD90 and CD7 in the RAG1+ population from differentiation days 16–24 showing the progressive loss of endothelial/haematopoietic progenitor markers over this time period (right). The position of cells falling within the RAG1+ gate that were used to generate accompanying contour plots (right) is indicated by the leftmost panels. The scale numbers for both the *x* and *y* axes are the exponents of $\log_{10}$[fluorescence]. The flow cytometry plots are representative examples drawn from $n=4$ experiments. The percentages of cells in each quadrant or associated with each gate are indicated. **b**, Summary data tracking the relationship between the expression of CD34 (top), and CD34, VECAD, CD90 and CD7 (bottom) over the same time period. The histograms show the mean ± s.e.m. for $n=5$ independent experiments. The connecting lines in the lower panel identify the mean at the specified time point for each of the indicated fractions. The expression of these markers is gradually lost, with the RAG1+ cells from cultures at day 24 being largely negative. **c**,**d**, Immunofluorescence analysis showing that the first RAG1+ cells co-express CAV1 and VECAD (**c**; differentiation day 16) and retain low levels of expression of SOX17 (**d**, white arrows; differentiation day 20). The cell nuclei are highlighted with 4,6-diamidino-2-phenylindole (DAPI). Scale bars, 20 (**c**) and 50 μm (**d**). **e**, Fluorescence images of live cultures over consecutive days showing the relationship between RAG1+ cells and endothelial structures (CD31). The white arrows indicate cells that have migrated away from the endothelium. Scale bars, 100 and 300 μm (magnifications of the white boxes, right).

Analysis of these same fractions cultured under conditions that supported T-lineage development (IL7/FLT3L) showed that the CD34+VECAD− fraction generated the highest proportion of CD4+CD8+ cells, whereas the CD34−VECAD− fraction failed to generate any DP cells (Fig. 8e). Given this latter population was nominally the most developmentally advanced, the absence of DP





cells prompted us reanalyse the same cultures at a later time point. This analysis revealed that the proportion of CD5+CD7+ cells was substantially lower in cultures derived from CD34−VECAD− cells relative to other factions (Fig. 8f,g). Instead, this last fraction generated CD161+ (KLRB1) cells, a finding that reflects our single-cell sequencing results, which suggested the presence of innate lymphoid

Fig. 8 | Relationship between the expression levels of RAG1 (GFP), the progenitor markers CD34, VECAD and potentiality. a, Flow cytometry analysis showing the relationship between the expression of RAG1+, CD34 and VECAD at the early stages of differentiation (day 19). **b–g**, Sort and re-culture experiments of RAG1+ cells expressing CD34 and VECAD. **b**, Representative flow cytometry gates used to isolate the indicated fractions. Sorted cells were seeded onto DLL4-OP9 stromal cells in the presence of DAPT, IL3 and SCF. The blood cells were analysed by flow cytometry for expression of CD10, CD19 and CD14 after 16 d. **c**, Summary of $n=4$ sort and re-culture experiments, in which stromal co-cultures containing DAPT, SCF and IL3 were analysed for CD10+ and CD19+ (B lineage), and CD14+ (myeloid lineage) cells. **d**, Bright-field images of stromal co-cultures illustrating that cells derived from the CD34+VECAD+ fraction were substantially more proliferative than the CD34−VECAD− population. Scale bars, 200 μm. **e–g**. Flow cytometry analysis of OP9-DLL4 co-cultures derived from the fractions depicted in **b** following culture in media supplemented with IL7 and FLT3L for expression of CD4 and CD8 (**e**; day 19), and CD5, CD7 and CD161 (**f**; day 27). **g**, Summary of the data from **f**. V, VECAD; 34, CD34; $n=2$ (CD34+VECAD+), 3 (CD34+VECAD−) and 5 (CD34−VECAD−). The mean and s.e.m. are shown. **h**, Summary of RAG1+ cell development in haematopoietic organoids. Conventionally, RAG1+ cells are thought to arise from early haematopoietic progenitors following their emergence from haemogenic endothelium (orange). This study shows that RAG1 expression can be activated before the acquisition of the T-cell progenitor surface markers CD7 and CD5. Early RAG1+ cells are multipotent and capable of generating myeloid, erythroid and lymphoid lineages. The flow cytometry plots are representative examples drawn from multiple experiments; $n=4$ (**a**,**b**), 2 (**e**) and 3 (**f**); the percentages of cells in each quadrant or associated with each gate are indicated.

cells (NKT) in early RAG1+ populations. Overall, these experiments are consistent with a differentiation hierarchy in which the earliest RAG1+ cells possess multilineage differentiation potential—a potential that is progressively lost as cells downregulate the expression of progenitor markers.

Discussion

We used haematopoietic organoid cultures to directly observe the birth of human adaptive immune cells in vitro. In this system, RAG1 expression occurs within DLL4+SOX17+ haemogenic endothelium, which is also marked by CD31, VECAD, CD90 and the endothelial gene *CAV1*. Newly emerging RAG1+ cells are multipotent and able to give rise to myeloid, erythroid, B and/or T-lymphocyte progenitors (Fig. 8h). Our observations draw together several threads of evidence regarding the origins and potentiality of early multipotential lymphoid progenitors across diverse vertebrate systems.

The upregulation of RAG1 during the endothelial–haematopoietic transition mirrors findings in zebrafish showing that non-HSC-derived T-cell progenitors emerge directly from arterial endothelium[2]. This study indicated that T-cell progenitors were produced along the length of the aorta and that such cells developed in conjunction with, but separately from, HSCs. Analysis of *HOXA* gene expression in our organoids indeed suggests that although endothelial/haematopoietic progenitor fractions share many characteristics of AGM, they lack the expression of key *HOXA* genes that mark bona fide AGM.

Comparisons of our observations with those from mouse studies are aided by a well-characterized Rag1:GFP reporter mouse strain[44,45]. Using this strain, early Rag1+ cells have been identified in the yolk sac, fetal liver and thymic rudiment at developmental stages, which is suggestive of a pedigree independent of HSCs[3,4,17,41]. Similar to our findings, these early Rag1+ cells expressed endothelial/haematopoietic progenitor markers, including CD31 and VECAD, and displayed multilineage differentiation potential[3,4,41].

We speculate that the RAG1+ cells observed in our system are analogous to the T-cell progenitors observed in the zebrafish, which emerge directly from aortic haemogenic endothelium[2], and possibly those cultured from the AGM regions of human embryos at Carnegie stage 12–14 (ref. [46]). Such cells would be well placed to contribute to the pool of non-HSC-derived Rag1+ progenitors observed in thymic rudiments[3], potentially providing a source of T cells that are critical for proper thymic maturation[47,48]. In conclusion, we describe an organoid culture system, independent of exogenous stromal cells, which has enabled us to observe the direct emergence of cells with T-cell potential from haemogenic endothelium. This system provides an accessible platform for further detailed study of the ontogeny of the earliest human adaptive immune cells.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41556-019-0445-8.

Received: 2 April 2019; Accepted: 27 November 2019;
Published online: 6 January 2020

References

1. Lin, Y., Yoder, M. C. & Yoshimoto, M. Lymphoid progenitor emergence in the murine embryo and yolk sac precedes stem cell detection. *Stem Cells Dev.* **23**, 1168–1177 (2014).
2. Tian, Y. et al. The first wave of T lymphopoiesis in zebrafish arises from aorta endothelium independent of hematopoietic stem cells. *J. Exp. Med.* **214**, 3347–3360 (2017).
3. Luis, T. C. et al. Initial seeding of the embryonic thymus by immune-restricted lympho-myeloid progenitors. *Nat. Immunol.* **17**, 1424–1435 (2016).
4. Boiers, C. et al. Lymphomyeloid contribution of an immune-restricted progenitor emerging prior to definitive hematopoietic stem cells. *Cell Stem Cell* **13**, 535–548 (2013).
5. Boiers, C. et al. A human IPS model implicates embryonic B-myeloid fate restriction as developmental susceptibility to B acute lymphoblastic leukemia-associated ETV6-RUNX1. *Dev. Cell* **44**, 362–377 (2018).
6. Ivanovs, A. et al. Highly potent human hematopoietic stem cells first emerge in the intraembryonic aorta-gonad-mesonephros region. *J. Exp. Med.* **208**, 2417–2427 (2011).
7. Bertrand, J. Y. & Traver, D. Hematopoietic cell development in the zebrafish embryo. *Curr. Opin. Hematol.* **16**, 243–248 (2009).
8. Chen, A. T. & Zon, L. I. Zebrafish blood stem cells. *J. Cell. Biochem.* **108**, 35–42 (2009).
9. Medvinsky, A. & Dzierzak, E. Definitive hematopoiesis is autonomously initiated by the AGM region. *Cell* **86**, 897–906 (1996).
10. Yoshimoto, M. et al. Autonomous murine T-cell progenitor production in the extra-embryonic yolk sac before HSC emergence. *Blood* **119**, 5706–5714 (2012).
11. Yoshimoto, M. et al. Embryonic day 9 yolk sac and intra-embryonic hemogenic endothelium independently generate a B-1 and marginal zone progenitor lacking B-2 potential. *Proc. Natl Acad. Sci. USA* **108**, 1468–1473 (2011).
12. Godin, I., Dieterlen-Lièvre, F. & Cumano, A. Emergence of multipotent hemopoietic cells in the yolk sac and paraaortic splanchnopleura in mouse embryos, beginning at 8.5 days postcoitus. *Proc. Natl Acad. Sci. USA* **92**, 773–777 (1995).
13. Nishikawa, S. I. et al. In vitro generation of lymphohematopoietic cells from endothelial cells purified from murine embryos. *Immunity* **8**, 761–769 (1998).
14. Kawamoto, H., Ikawa, T., Ohmura, K., Fujimoto, S. & Katsura, Y. T cell progenitors emerge earlier than B cell progenitors in the murine fetal liver. *Immunity* **12**, 441–450 (2000).
15. Ikawa, T. et al. Identification of the earliest prethymic T-cell progenitors in murine fetal blood. *Blood* **103**, 530–537 (2004).
16. Benz, C. & Bleul, C. C. A multipotent precursor in the thymus maps to the branching point of the T versus B lineage decision. *J. Exp. Med.* **202**, 21–31 (2005).

17. Yokota, T. et al. Tracing the first waves of lymphopoiesis in mice. *Development* **133**, 2041–2051 (2006).
18. Mombaerts, P. et al. RAG-1-deficient mice have no mature B and T lymphocytes. *Cell* **68**, 869–877 (1992).
19. Sobacchi, C., Marrella, V., Rucci, F., Vezzoni, P. & Villa, A. RAG-dependent primary immunodeficiencies. *Hum. Mutat.* **27**, 1174–1184 (2006).
20. Ivanovs, A. et al. Human haematopoietic stem cell development: from the embryo to the dish. *Development* **144**, 2323–2337 (2017).
21. Tavian, M., Biasch, K., Sinka, L., Vallet, J. & Peault, B. Embryonic origin of human hematopoiesis. *Int. J. Dev. Biol.* **54**, 1061–1065 (2010).
22. Ditadi, A. et al. Human definitive haemogenic endothelium and arterial vascular endothelium represent distinct lineages. *Nat. Cell Biol.* **17**, 580–591 (2015).
23. Sturgeon, C. M., Ditadi, A., Awong, G., Kennedy, M. & Keller, G. Wnt signaling controls the specification of definitive and primitive hematopoiesis from human pluripotent stem cells. *Nat. Biotechnol.* **32**, 554–561 (2014).
24. Kennedy, M. et al. T lymphocyte potential marks the emergence of definitive hematopoietic progenitors in human pluripotent stem cell differentiation cultures. *Cell Rep.* **2**, 1722–1735 (2012).
25. Mohtashami, M. et al. Direct comparison of Dll1- and Dll4-mediated Notch activation levels shows differential lymphomyeloid lineage commitment outcomes. *J. Immunol.* **185**, 867–876 (2010).
26. Carpenter, L. et al. Human induced pluripotent stem cells are capable of B-cell lymphopoiesis. *Blood* **117**, 4008–4011 (2011).
27. Ng, E. S. et al. Differentiation of human embryonic stem cells to HOXA$^+$ hemogenic vasculature that resembles the aorta-gonad-mesonephros. *Nat. Biotechnol.* **34**, 1168–1179 (2016).
28. Bertrand, J. Y., Cisson, J. L., Stachura, D. L. & Traver, D. Notch signaling distinguishes 2 waves of definitive hematopoiesis in the zebrafish embryo. *Blood* **115**, 2777–2783 (2010).
29. Hadland, B. K. et al. A requirement for Notch1 distinguishes 2 phases of definitive hematopoiesis during development. *Blood* **104**, 3097–3105 (2004).
30. Kumano, K. et al. Notch1 but not Notch2 is essential for generating hematopoietic stem cells from endothelial cells. *Immunity* **18**, 699–711 (2003).
31. Schulz, T. C. et al. A scalable system for production of functional pancreatic progenitors from human embryonic stem cells. *PLoS ONE* **7**, e37004 (2012).
32. Taoudi, S. et al. Extensive hematopoietic stem cell generation in the AGM region via maturation of VE-cadherin$^+$CD45$^+$ pre-definitive HSCs. *Cell Stem Cell* **3**, 99–108 (2008).
33. Skelton, R. J. et al. CD13 and ROR2 permit isolation of highly enriched cardiac mesoderm from differentiating human embryonic stem cells. *Stem Cell Rep.* **6**, 95–108 (2016).
34. Gama-Norton, L. et al. Notch signal strength controls cell fate in the haemogenic endothelium. *Nat. Commun.* **6**, 8510 (2015).
35. Park, S. H. et al. HLA-DR expression in human fetal thymocytes. *Hum. Immunol.* **33**, 294–298 (1992).
36. Melichar, H. J., Ross, J. O., Taylor, K. T. & Robey, E. A. Stable interactions and sustained TCR signaling characterize thymocyte–thymocyte interactions that support negative selection. *J. Immunol.* **194**, 1057–1061 (2015).
37. Lancaster, J. N., Li, Y. & Ehrlich, L. I. R. Chemokine-mediated choreography of thymocyte development and selection. *Trends Immunol.* **39**, 86–98 (2018).
38. Zlotoff, D. A. et al. CCR7 and CCR9 together recruit hematopoietic progenitors to the adult thymus. *Blood* **115**, 1897–1905 (2010).
39. Plotkin, J., Prockop, S. E., Lepique, A. & Petrie, H. T. Critical role for CXCR4 signaling in progenitor localization and T cell differentiation in the postnatal thymus. *J. Immunol.* **171**, 4521–4527 (2003).
40. Koenen, P. et al. Mutually exclusive regulation of T cell survival by IL-7R and antigen receptor-induced signals. *Nat. Commun.* **4**, 1735 (2013).
41. Yokota, T. et al. Unique properties of fetal lymphoid progenitors identified according to RAG1 gene expression. *Immunity* **19**, 365–375 (2003).
42. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
43. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
44. Kuwata, N., Igarashi, H., Ohmura, T., Aizawa, S. & Sakaguchi, N. Cutting edge: absence of expression of RAG1 in peritoneal B-1 cells detected by knocking into RAG1 locus with green fluorescent protein gene. *J. Immunol.* **163**, 6355–6359 (1999).
45. Igarashi, H. et al. Localization of recombination activating gene 1/green fluorescent protein (RAG1/GFP) expression in secondary lymphoid organs after immunization with T-dependent antigens in *rag1/gfp* knockin mice. *Blood* **97**, 2680–2687 (2001).
46. Tavian, M., Robin, C., Coulombel, L. & Peault, B. The human embryo, but not its yolk sac, generates lympho-myeloid stem cells: mapping multipotent hematopoietic cell fate in intraembryonic mesoderm. *Immunity* **15**, 487–495 (2001).
47. Klug, D. B. et al. Interdependence of cortical thymic epithelial cell differentiation and T-lineage commitment. *Proc. Natl Acad. Sci. USA* **95**, 11822–11827 (1998).
48. Hess, I. & Boehm, T. Intravital imaging of thymopoiesis reveals dynamic lympho-epithelial interactions. *Immunity* **36**, 298–309 (2012).

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

Methods

PSC culture. The culturing of undifferentiated PSCs was performed as described[49,50]. The reagents for PSC culture and differentiation medium were purchased from Thermofisher, unless otherwise specified. Briefly, hESC and iPSC cell lines were grown in PSC media consisting of DMEM-F12, 20% knock-out serum replacement, 1×non-essential amino acids, 1×GlutaMAX, 0.11 mM β-mercaptoethanol and FGF2 (Peprotech; 10 ng ml^{-1} for hESCs and 50 ng ml^{-1} for iPSCs) in the presence of mouse embryonic fibroblasts. Once a confluency of approximately 90% was reached, the PSCs were passaged using EDTA dissociation buffer (PBS) supplemented with 100 mM NaCl and 0.5 mM EDTA. The cells were then transferred to new flasks that had been pre-seeded with mouse embryonic fibroblasts at approximately 15,000 cells cm^{-2}.

Generation of RAG1^{GFP} PSC lines. Cells were prepared for electroporation using a Neon transfection system as described previously[50]. The RAG1–GFP targeting construct (see Extended Data Fig. 1), in combination with TALEN pairs designed to cut the *RAG1* gene at a site immediately 3′ of the *RAG1* start codon in exon 2 were electroporated into H9 hESCs[51] and RM3.5 iPSCs[50]. Puromycin (1 µg ml^{-1}) was added to the culture medium 2 d after electroporation and selection was continued for 2 d, after which the medium was changed to PSC medium as detailed above. At 10 d post electroporation, individual colonies were picked, expanded and subsequently screened for the correct targeting events using the PCR strategy outlined in Extended Data Fig. 1a. The sequence of the PCR primers is listed in Supplementary Table 1.

PSC clones that had sequences encoding GFP inserted into one allele and an indel-free second allele (as determined by DNA sequencing of the relevant PCR product) were chosen for further analysis. Two clones representing the H9 hESC and RM3.5 iPSC parental lines were expanded in preparation for removal of the PGKpuro selectable marker cassette. This cassette was removed by electroporating cells representing each clone with a CRE expression plasmid[52]. Individual cells were sorted into each well of a 96-well tray using the single-cell deposition function of a FACSAria flow cytometer 3 d following electroporation. The single-cell-derived clones were then screened for excision of the drug-resistance cassette using a primer set flanking the cassette, as indicated in Extended Data Fig. 1a and Supplementary Table 1. As a final check for cell lines designated as RAG1$^{GFP/w}$, sequences representing the wild-type allele were amplified by PCR and subjected to analysis through Sanger sequencing.

Cell lines. The origin of the parental RM3.5 iPSC and H9 hESC cell lines used in this study have been described previously[50,51]. The mouse embryonic fibroblasts used for PSC culture were isolated as previously described[49] from mouse embryos as per the mouse animal ethics approval A774 (Murdoch Children's Research Institute).

Haematopoietic differentiation. The day before initiation of differentiation, PSCs were passaged to new flasks previously coated with feeder medium (DMEM), 10% FCS, 1×GlutaMAX and 1×penicillin–streptomycin. On the day of differentiation, the cells were dissociated using EDTA dissociation buffer and resuspended in organoid differentiation medium (ODM) containing STAGE I supplements. The cells were then passed through a cell-strainer cap (pore size of 40 µm) to remove large cell clumps and immediately transferred to non-tissue-culture treated 6-cm dishes (4–5 ml per dish). The 6-cm dishes were then placed on a Raytech rotating platform inside a 37 °C incubator. The swirler cultures were rotated at 60 r.p.m.

At day 1 of differentiation, medium containing the embryoid bodies was transferred to a Falcon tube and the embryoid bodies were allowed to settle to the bottom of the tube (5–10 min). The medium was carefully aspirated from the tube and replaced with ODM medium containing STAGE II supplements. Similarly, media changes were performed at days 3, 4 and 6 of differentiation. ODM medium containing STAGE III and STAGE IV supplements was used for differentiation days 2–4 and 4–8, respectively. The formulation of ODM and the relevant supplements are detailed below.

The ODM medium was prepared by mixing 0.1% (or 0.25% for hESCs) Recombumin alpha (recombinant human albumin; Albumedix), 0.1% methylcellulose (Sigma-Aldrich), 0.1% polyvinyl alcohol (Sigma-Aldrich), 1×GlutaMAX, 1×ascorbic acid-2-phosphate (Sigma-Aldrich), ITSE AF blood-free cell culture media supplement (50 µg ml^{-1}; InVitria), linoleic and linolenic acid (100 ng ml^{-1}; Sigma-Aldrich), synthetic cholesterol (2.2 µg ml^{-1}; Sigma-Aldrich), 2-mercaptoethanol (22 nM; 55 nM for hESCs) and protein-free hybridoma mix II (4%) in IMDM/F12 media. Unless otherwise specified, growth factors were purchased from Peprotech.

The STAGE I supplements were: rock inhibitor (10 µM; 20 µM for the hESC cell line; Y-27632, Stem Cell Technologies), CHIR99021 (0.5 µM; Tocris), activin A (10 ng ml^{-1}; R&D Systems), BMP4 (20–40 ng ml^{-1}; R&D Systems), SCF (20 ng ml^{-1}), VEGF (20 ng ml^{-1}) and bFGF (5–10 ng ml^{-1}).

The STAGE II supplements were: CHIR99021 (0.5 µM), activin A (10 ng ml^{-1}), BMP4 (20 ng ml^{-1}), SCF (20 ng ml^{-1}), VEGF (20 ng ml^{-1}) and bFGF (10 ng ml^{-1}).

The STAGE III supplements were: CHIR99021 (3 µM), SB-431542 (3 µM; Cayman Chemical), BMP4 (20 ng ml^{-1}), SCF (20 ng ml^{-1}), VEGF (20 ng ml^{-1}) and bFGF (10 ng ml^{-1}).

The STAGE IV supplements were: BMP4 (20 ng ml^{-1}), VEGF (50 ng ml^{-1}), SCF (50 ng ml^{-1}), IGFII (20 ng ml^{-1}) and bFGF (10 ng ml^{-1}).

T/NK-cell differentiation on ALI cultures. Embryoid bodies were transferred to vitronectin-coated tissue-culture-treated ALI membranes (Corning transwell culture plates, CLS3450) at day 8. The ODM medium containing the appropriate cytokines, which was placed in the chamber below the ALI membrane, was changed every 3 d. Cytokines were added to the cultures during this phase of differentiation as listed below.

For days 8–17, VEGF (50 ng ml^{-1}), SCF (100 ng ml^{-1}), bFGF (10 ng ml^{-1}) and IL7 (20 ng ml^{-1}) were added. FLT3L (10 ng ml^{-1}) and IL2 (10 ng ml^{-1}; Peprotech) were sometimes included, although the effect of this variation was not explicitly assessed.

For days 17–26, VEGF (50 ng ml^{-1}), SCF (20 ng ml^{-1}), bFGF (10 ng ml^{-1}) and IL7 (20 ng ml^{-1}) were added.

From day 26, the medium was supplemented with only VEGF (50 ng ml^{-1}) and IL7 (20 ng ml^{-1}).

For NK differentiation, experiments were performed similar to T-cell differentiation except that IL2 (10 ng ml^{-1}) and IL15 (10 ng ml^{-1}) were added from the initiation of ALI culture.

B-cell differentiation on ALI cultures. For B-cell induction on ALI cultures, differentiation day 8 embryoid bodies were transferred onto vitronectin-coated transwell membranes in contact with ODM medium supplemented with VEGF (50 ng ml^{-1}), SCF (100 ng ml^{-1}), bFGF (10 ng ml^{-1}), FLT3L (10 ng ml^{-1}) and IL3 (10 ng ml^{-1}). The medium was changed every 3 d. Once haematopoietic cells were observed (visual inspection) at around day 15–20 of differentiation, the medium was changed to ODM supplemented with VEGF (50 ng ml^{-1}), SCF (100 ng ml^{-1}), bFGF (10 ng ml^{-1}), IL7 (10 ng ml^{-1}), IL3 (10 ng ml^{-1}), IL6 (10 ng ml^{-1}), G-CSF (10 ng ml^{-1}) and DAPT (10 µM; Santa Cruz Biotechnology). The medium was changed every 3 d.

T/B-cell differentiation on stromal cells. Cell populations purified using FACS, as indicated, were seeded onto pre-prepared confluent layers of OP9-DLL4 (refs. [25,53]; T-cell experiments) or OP9 (B-cell experiments) stromal cells grown on 12-well plates. Approximately 1×10^3 cells were seeded per well. The cells were cultured in alpha-MEM media containing 10% FCS, penicillin–streptomycin, GlutaMax and 2-mercaptoethanol (0.1 mM), supplemented with SCF (50 ng ml^{-1}; first week only) and IL7 (10 ng ml^{-1}; sometimes 10 ng ml^{-1} FLT3L; T-cell differentiation on OP9-DLL4) or SCF (50 ng ml^{-1}; sometimes 10 ng ml^{-1} IL3 and 10 ng ml^{-1} G-CSF for B-cell differentiation on OP9). For the data presented in Fig. 8 relating to B- and myeloid-lineage differentiations, cells were cultured on OP9-DLL stromal cells in media supplemented with 10 µM DAPT, 50 ng ml^{-1} SCF and 10 ng ml^{-1} IL3. Haematopoietic cells were transferred onto fresh layers of stromal cells every week.

Cell harvesting, flow cytometry analysis and sorting. The methods used for harvesting cells in preparation for flow cytometry were dependent on each specific culture condition. To harvest cells from embryoid bodies at days 4–8, the embryoid bodies were collected by allowing them to settle in a 15-ml Falcon tube, as described earlier. Following a single wash with PBS, the embryoid bodies were treated with 2 ml TrypleSelect for 5–10 min and then disaggregated by vigorous pipetting. The cells were washed in FACS wash (PBS with 5% FCS) before antibody labelling. To harvest cells from ALI cultures, 1 mg ml^{-1} collagenase type I (Sigma-Aldrich) was added to the top side of the membrane and the cultures were returned to the 37 °C incubator for at least 1 h. Following this, the cells were transferred to a 15-ml Falcon tube, disaggregated and washed as above. To harvest haematopoietic cells from stromal co-cultures, only physical agitation (pipetting up and down) was used to dislodge non-adherent cells, which were then transferred to a 15-ml Falcon tube and processed as above.

The cell mixtures were passed through the cell-strainer cap of a FACS tube to obtain a single-cell suspension. The samples were centrifuged for 3 min at 1,500 r.p.m., the supernatant was removed and the cells were resuspended in FACS wash containing the indicated combinations of antibodies (see Supplementary Table 2 for flow cytometry antibodies). Antibody labelling proceeded on ice for 15 min. Following this incubation, the cells were washed twice with 2 ml FACS wash. The cells were resuspended in FACS wash containing 1 µg ml^{-1} propidium iodide before analysis or sorting to identify the live cells.

Flow cytometry analysis was performed using the BD Fortessa analyzer. Cell sorting was performed using the BD FACSAria FUSION and BD InFlux.

Neonatal thymus tissue collection. Neonatal thymus tissue was obtained from The Royal Children's Hospital from a six-week-old female paediatric patient with trisomy 21. This child was diagnosed with a large perimembranous ventricular septal defect, atrial septal defect and patent ductus arteriosus and underwent cardiac surgery (atrial septal defect, ventricular septal defect and patent ductus arteriosus closure). Tissue collection for research purposes was obtained under the human ethics approval HREC 24131H following informed consent by a parent or guardian. The thymus tissue was mechanically disrupted to release thymocytes. Briefly, the thymus tissue was cut into pieces of approximately 0.5 cm^3 and then, using the plunger of a 10-ml disposable syringe, pressed against the membrane of a 40-µm cell strainer sitting in a sterile 6-cm tissue-culture plate. The dislodged cells were flushed through the membrane using cold FACS wash. The cell solution was then passed through the cell-strainer cap of a FACS tube to obtain the single-cell suspension. Antibody labelling to identify the indicated cell populations was performed as described above.

Gene expression analysis by qPCR. RNA was extracted from purified populations using a Isolate II RNA micro kit and the method suggested by the manufacturer (cat. no. BIO-52075, Bioline). Complementary DNA was prepared using a Tetro cDNA synthesis kit (cat. no. BIO-65042, Bioline) according to the instructions provided with the kit. Quantitative PCR analysis was conducted on an ABI real-time PCR (Applied Biosystems) machine. Reaction samples (20 µl) containing cDNA, Taqman universal PCR master mix and Taqman assay probes were prepared in a MicroAmp 96-well optical reaction plate. Data were analysed using ABI 7300 software (Version 1.4.1). *RAG1* mRNA was detected using the Taqman probe hs00172121_m1 RAG1.

Colony-forming assays. The colony-forming assays were performed as described previously[27].

Imaging. Images were captured using a Zeiss Axiovert 200 microscope, Zeiss LSM 780 confocal microscope and Dragonfly spinning disk confocal microscope, and analysed using ZEN blue (Zeiss, version 2.1), ImageJ (Version 1.51) and Imaris software (Version 9.2.1). The images were then exported as JPEG or PNG files and labelled using Adobe illustrator (Version 23.1).

Immunofluorescence staining. Organoids were washed once with PBS to remove excess media and fixed in PBS with 4% paraformaldehyde (Santa Cruz Biotechnology) for 10 min at room temperature. Excess paraformaldehyde solution was then removed by washing the organoids three times in PBS. The fixed organoids were stored in PBS at 4 °C until required. For antibody labelling, the fixed organoids were blocked in 10% donkey serum in PBS with 0.3% Triton X-100 (blocking buffer) for at least 1 h at room temperature before incubation with the primary antibodies (listed in Supplementary Table 2), which were also prepared in blocking buffer. The primary antibodies used for the analysis of organoids were: mouse anti-human CD31–APC, mouse anti-human VECAD–AF647, rabbit anti-human caviolin-1 (1:200 dilution; Abcam, cat. no. ab18199), chicken anti-GFP (1:500 dilution; Abcam, cat. no. ab13970), rabbit anti-human GATA3 (1:200 dilution; Cell Signaling Technology, cat. no. 5852S) and goat anti-human SOX17 (1:100 dilution; R&D Systems, cat. no. AF1924). Nuclei were identified by labelling organoids with 1 µg ml^{-1} DAPI (Thermofisher). The organoids were incubated with primary antibodies overnight at 4 °C, washed five times in blocking buffer and then incubated with the species-appropriate fluorophore-conjugated secondary antibodies—donkey anti-rabbit AF-568 (1:1,000 dilution; Life Technologies, cat. no. A10042), donkey anti-chicken AF-488 (1:1,000 dilution; Jackson Immuno, cat. no. 703-545-155) and donkey anti-goat AF647 (1:1,000 dilution; Life Technologies, cat. no. A21447). The labelled organoids were mounted on a MatTek glass-bottom dish and confocal microscopy was performed using an inverted Zeiss LSM 780 microscope.

Bulk RNAseq analysis. RNA extraction of FACS-purified populations was performed using TRIzol reagent (Thermofisher) according to the manufacturer's guidelines. The RNA samples were processed, quality controlled and sequenced by the Australian Genome Research Facility (AGRF). Sequencing of all samples was performed using a Nextseq500 (Illumina) instrument. Between 20 and 30 × 10^6 75-bp paired-end reads were obtained per sample. Mapping of the reads to the reference human genome (Homosapiens.GRCh38.91), initial quality control analysis and the generation of the raw count table of genes for the samples were performed by the Centre for Stem Cell Systems at the University of Melbourne. Individual fastq files were aligned to the reference genome (GRCh38 assembly and ENSEMBL version 91) using Rsubread 1.20.6 with default settings in R version 3.2.2. The reads were quantified with featureCounts as part of the Rsubread 1.20.6 library. The RNAseq analysis on raw count table was performed using R packages. Briefly, the genes expressed at low levels (counts-per-million (CPM) below 0.5) across all samples were first filtered out. The CPM value was calculated using the cpm() function in edgeR[54,55]; only genes with CPM > 0.5 in at least one sample were kept for further analysis. Next, the filtered table of counts was used to create a DGElist object through the DGElist() function in EdgeR. The object was used to create multidimensional-scaling plots using the plotMDS() function in EdgeR. Highly variable genes were then identified by estimating the variance of each gene across samples and sorting the genes according to the variance value. The Heatmap.2() function was used for unsupervised hierarchical clustering of samples using the top-500 variable genes. The object was then TMM normalized using the calcNormFactors () function. The DGElist object was then prepared to test for differentially expressed genes via voom transformation using the voom() function of the limma package[56]. Genes whose expression was significantly different between the desired sets of samples were identified using the lmfit(), contrasts.fit (), eBayes() and decideTests () functions (in their default settings) from the limma package. The RPKM values of each dataset were created using the rpkm() function in edgeR using the TMM-normalized DGElist object.

Mapping GFP, RAG1 and RAG2 reads. To count the reads that were mapped to GFP sequences, the sequence was added as an extra chromosome to the genome reference FASTA file and its annotation was incorporated in the GTF annotation file. All of the files were uploaded on to the GALAXY server (usegalaxy.org). STAR aligner[57] was used for mapping and counting the number of reads per gene.

Analysis of scRNA-seq data. Two independent experimental sets of purified iPSC-derived RAG1:GFP+ cells were submitted to the AGRF for scRNA-seq using the 10x Genomics Chromium system. The sequencing and data generation were performed using the AGRF Illumina bcl2fastq pipeline version 2.20.0.422 through the generation of 100-bp paired-end reads and sequencing using the Illumina HiSeq platform. The 10x single-cell software (Cell Ranger) was used to perform the secondary analysis, including generation of Cell Ranger mkfastq and Cell Ranger count on each of the samples. This process produced several output files per sample, including filtered feature-barcode matrices MEX (filtered_gene_bc_matrices/hg38/matrix.mtx). This file was set as an input for further analysis in R using the Seurat package[58]. Quality control of the data and comprehensive analysis of the samples were performed on the basis of the tutorial guidelines provided by Satijilab (https://satijalab.org/seurat/get_started.html). Due to the 'dropout' issue often observed in the scRNA-seq data, we also used the data diffusion method MAGIC recently developed by the Krishnaswamy lab[42] where indicated. For this purpose, after normalization of data in Seurat, we used the magic() function in its default format and included all genes for data diffusion. The MAGIC-imputed data were then subjected to further analysis in Seurat, similar to the original dataset.

The Monocole package[43] in R was used for single-cell trajectory analysis. Seurat objects were converted into a CellDataSet object. Low-quality cells were filtered out using the same criteria as specified in Seurat. Cells were classified into four groups: endothelial, haematopoietic, T cell and NKT cell on the basis of the markers indicated in Fig. 5f. Cells marked as 'Ambiguous' or 'Unknown' were discarded. The remaining cells were used to generate a single-cell trajectory using the guidelines provided by the Trapnell laboratory (http://cole-trapnell-lab.github.io/monocle-release/docs_mobile/).

Statistics and reproducibility. For all data, the means, s.d. and s.e.m. were calculated using either Microsoft Excel (version 16.30) or GraphPad Prism 8 (version 8.2.1 (279)). No hypotheses that required statistical tests have been argued in this manuscript. The reproducibility of specific numerical data is encapsulated in the number of experimental replicates, as indicated in the figure legends and Reporting Summary.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All scRNA-seq and bulk RNA-seq data used in this study have been deposited in the Gene Expression Omnibus (GEO) database repository and are available under accession numbers GSE124172 (scRNA-seq) and GSE124173 (bulk RNAseq). All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Code availability

All computational code used for the analysis in this manuscript is available from the corresponding author upon reasonable request.

References

49. Costa, M. et al. A method for genetic modification of human embryonic stem cells using electroporation. *Nat. Protoc.* **2**, 792–796 (2007).
50. Kao, T. et al. GAPTrap: a simple expression system for pluripotent stem cells and their derivatives. *Stem Cell Rep.* **7**, 518–526 (2016).
51. Thomson, J. A. et al. Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
52. Davis, R. P. et al. A protocol for removal of antibiotic resistance cassettes from human embryonic stem cells genetically modified by homologous recombination or transgenesis. *Nat. Protoc.* **3**, 1550–1558 (2008).
53. Mohtashami, M., Shah, D. K., Kianizad, K., Awong, G. & Zúñiga-Pflücker, J. C. Induction of T-cell development by Delta-like 4-expressing fibroblasts. *Int. Immunol.* **25**, 601–611 (2013).
54. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
55. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
56. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

57. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
58. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

Acknowledgements

We would like to thank the members of the MCRI Flow cytometry support team for their assistance with cell sorting and microscopy and S. Chen for her assistance with the preparation of the art work. This study was funded by the National Health and Medical Research Council of Australia through research fellowships awarded to A.G.E. (grant no. GNT1117596) and E.G.S (grant no. GNT1079004), project grants awarded to A.G.E. and E.G.S. (grant nos GNT1068866, GNT1129861 and GNT1138717) as well as A.P.C. and E.G.S (grant no. GNT1123277), the Australian Research Council Special Research Initiative in Stem Cells (Stem Cells Australia), The Children's Cancer Research Foundation (Australia) and the Stafford Fox Medical Research Foundation. Additional infrastructure funding to the Murdoch Children's Research Institute was provided by the Australian Government National Health and Medical Research Council Independent Research Institute Infrastructure Support Scheme and the Victorian Government's Operational Infrastructure Support Program.

Author contributions

Conceptualization: A.M., A.G.E., E.G.S.; data curation: T.C.; formal analysis: A.M., F.F.B., T.C.; funding acquisition: A.P.C., C.A.W., A.G.E., E.G.S.; investigation: A.M., F.F.B., S.V.K., J.V.S., C.A.W.; methodology: A.M., F.F.B., E.S.N.; project administration: A.M., A.G.E., E.G.S.; resources: A.P.C.; supervision: A.G.E., E.G.S.; visualization: A.M., F.F.B., A.G.E., E.G.S.; writing—original draft: A.M., E.G.S.; writing—review and editing: A.M., A.P.C., C.A.W., A.G.E., E.G.S.

Competing interests

The authors declare no competing interests.

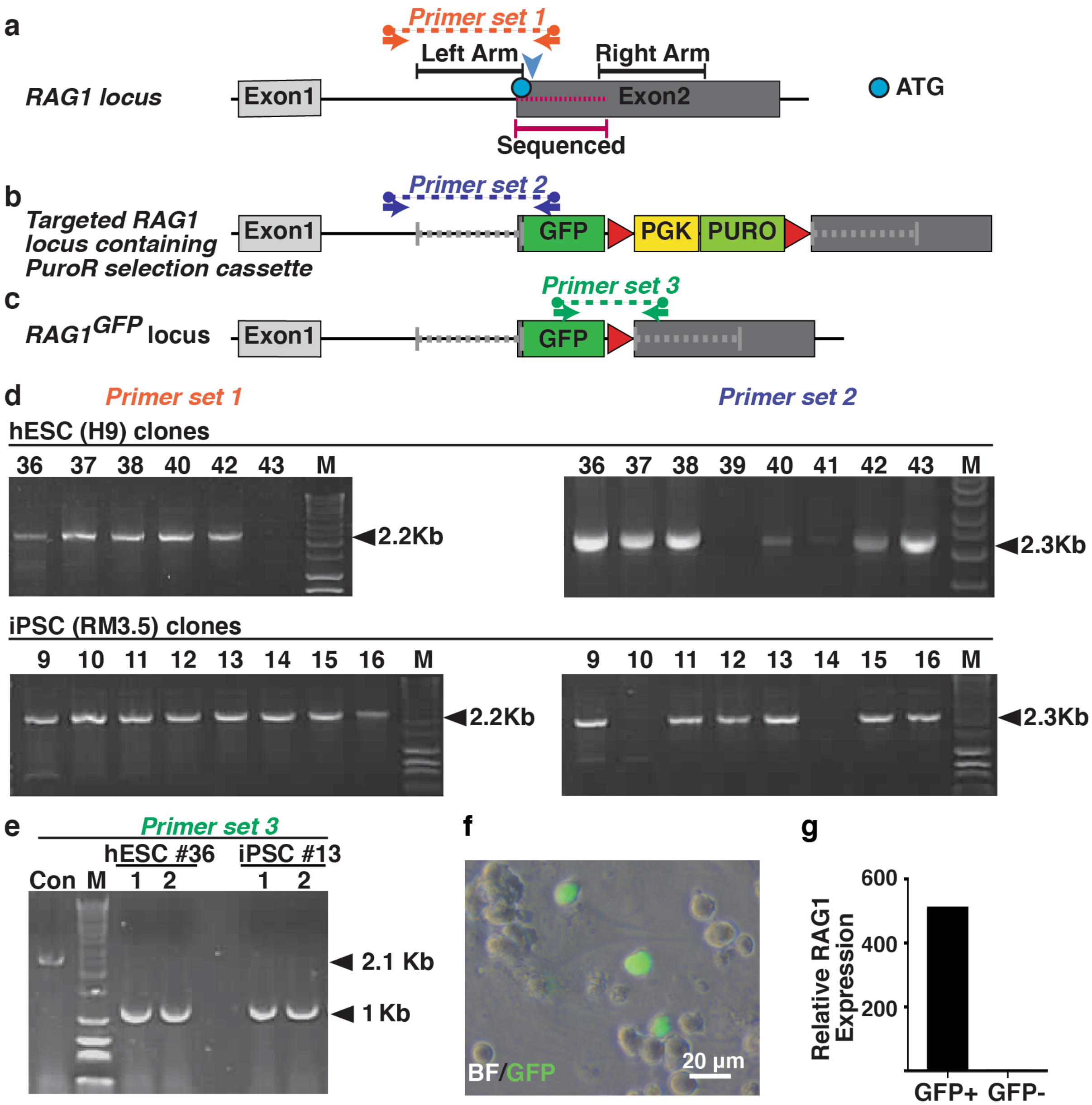
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41556-019-0445-8.

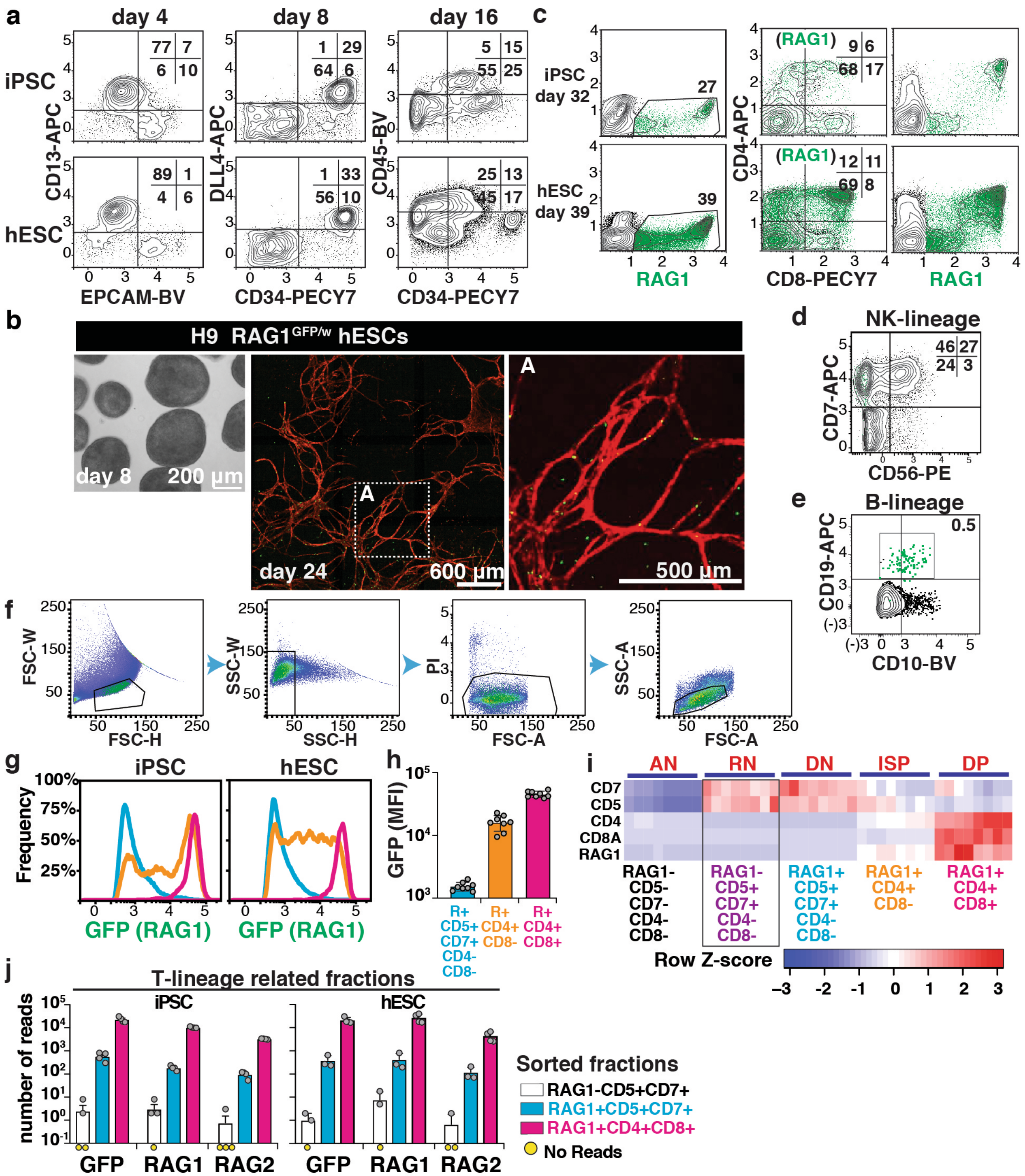
Supplementary information is available for this paper at https://doi.org/10.1038/s41556-019-0445-8.

Correspondence and requests for materials should be addressed to E.G.S.

Reprints and permissions information is available at www.nature.com/reprints.

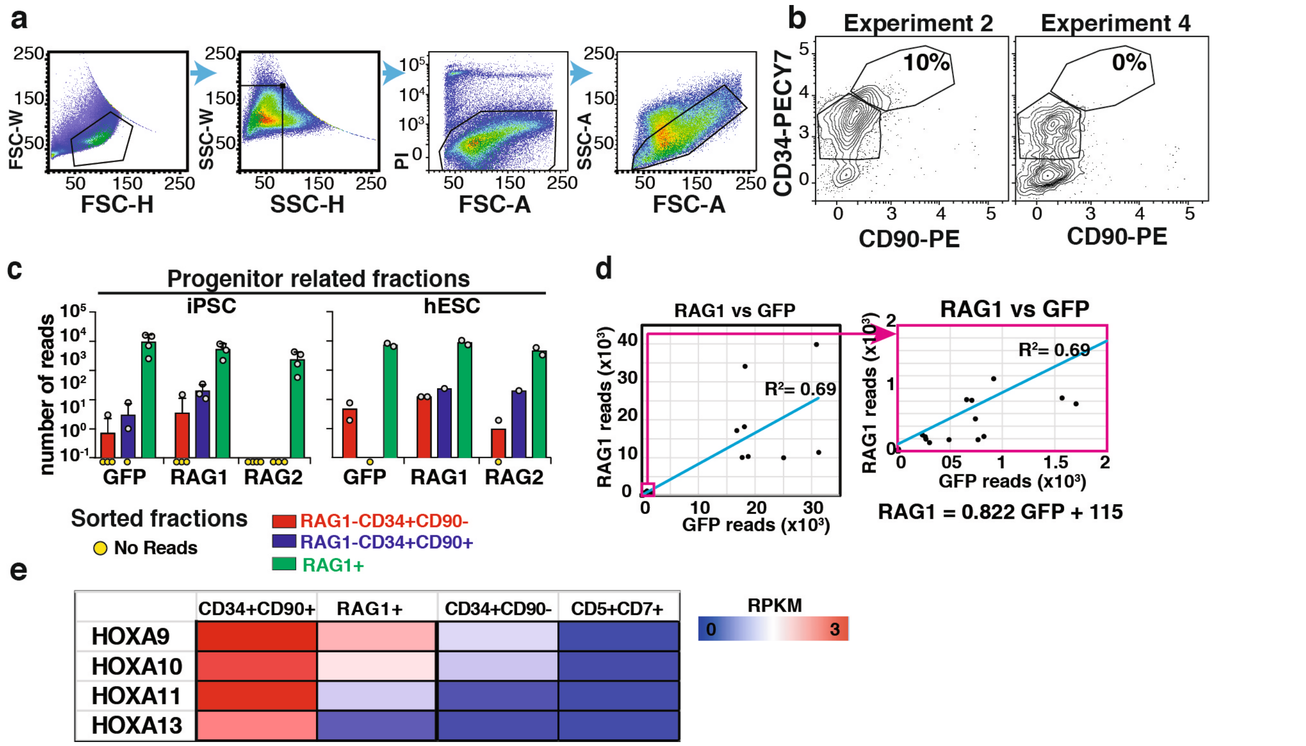


Extended Data Fig. 1 | Characterization of RAG1$^{GFP/w}$ PSC reporter lines. a, Schematic of the RAG1 locus showing the non-coding exon 1 and the coding exon 2. The cut site of RAG1 specific TALENs used to enhance gene targeting efficiency is indicated by a vertical blue arrowhead. The position of primer set 1 used to screen for clones that retained an intact RAG1 locus is indicated. The region of the native RAG1 locus in RAG1$^{GFP/w}$ PSC lines that was sequenced is indicated by the red dashed line. **b**, Schematic view of the RAG1 locus after insertion of a cassette encoding GFP and a loxP (red triangles) flanked selectable marker (PGK-Puro). Grey dashed lines denote sequences corresponding to those included in the targeting vector. **c**, Schematic view of the final targeted RAG1 locus. Primer set 3 was used to verify excision of sequences flanked by LoxP sites. **d**, Agarose gels showing the results of PCR analysis of genomic DNA representing individual H9 hESC and iPSC RM3.5 clones. As depicted in (**a**) and (**b**), primer set 1 amplified a DNA fragment of 2.2 Kb whereas primer set 2 amplified a DNA fragment of 2.3 Kb. Clone numbers corresponding to each PSC line are indicated. **e**, The loxP flanked selectable marker cassette was removed from the two heterozygous clones used in this study, H9 #36 and iPSC #13. Primer set 3 amplified a DNA fragment of 600 base pairs. control DNA derived from a targeted clone prior to the removal of the PGK-Puro cassette. **f**, Bright field-GFP fluorescence image showing individual GFP + human T-lymphoid progenitors. **g**, Histogram showing the results of real-time PCR indicating that RAG1 expression is restricted to the GFP + population. The Y axis shows the relative expression of RAG1 in arbitrary units from a single experiment. Results are derived from a single cell sorting experiment (n = 1). Scale bar = 20 μM.

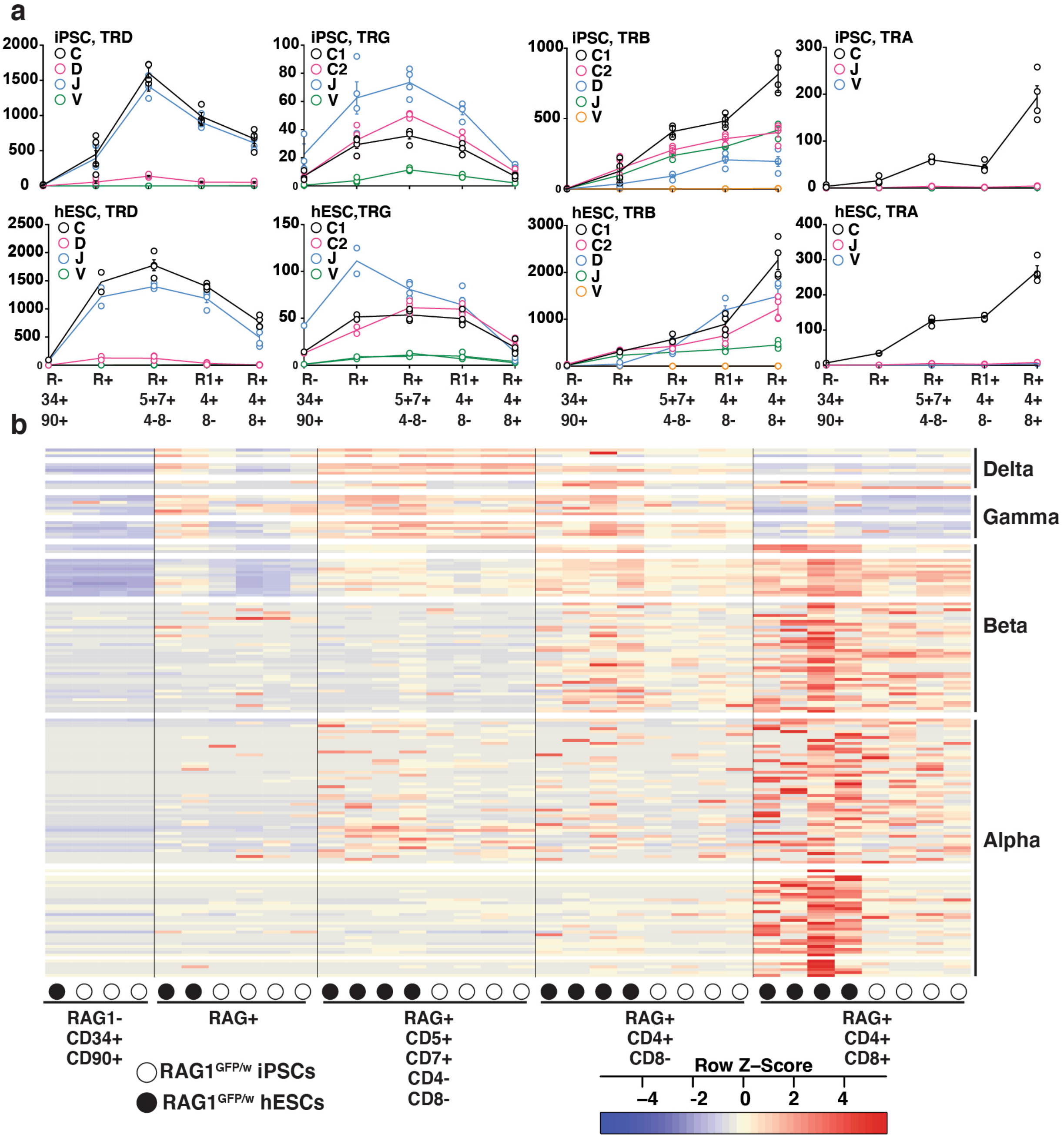


Extended Data Fig. 2 | See next page for caption.

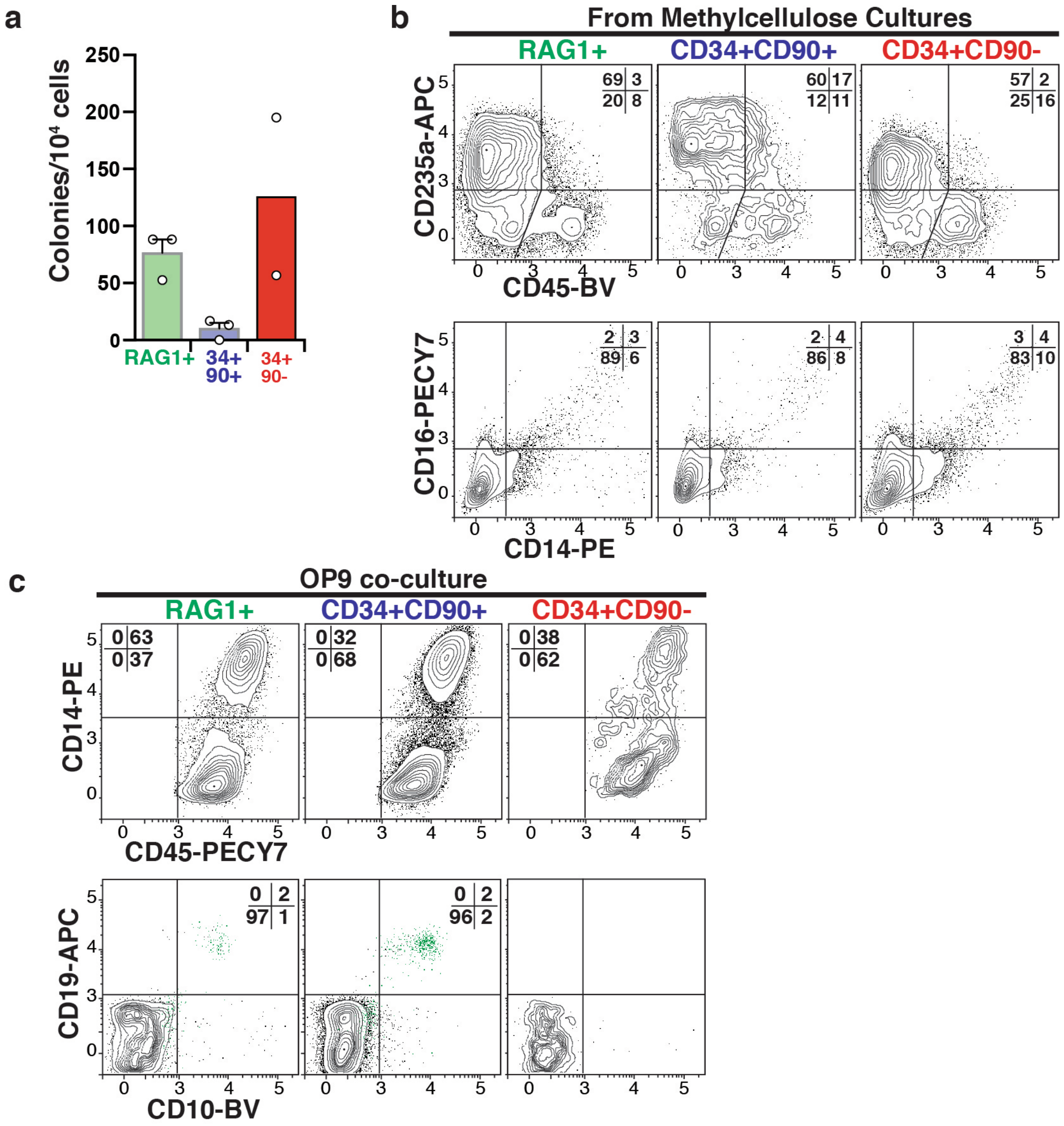
Extended Data Fig. 2 | Characterization of the haematopoietic organoid differentiation system. **a**, Flow cytometry data from iPSC and hESC lines showing the expression of the key indicated marker genes between differentiation days 4 and 16. **b**, Bright-field image of embryoid bodies at differentiation day 8 and fluorescent images of RAG1+(GFP+) cells within CD31+vascular networks of differentiation day 24 organoids. **c**, Flow cytometry plots showing the relationship between RAG1 expression levels and progression to the CD4+CD8+ stage. **d**, Flow cytometry analysis showing CD10+CD19+RAG1+B-cell progenitors in organoid cultures treated with the NOTCH signalling inhibitor, DAPT. **e**, Flow cytometry analysis quantifying the proportion of CD56+CD7+ cells (presumptive NK cells) in cultures supplemented with IL2 and IL15. **f**, Gating strategy used for isolation of cell populations profiled in Fig. 3. **g**, Relationship between GFP expression levels and specific T-cell differentiation stages. The frequency of cells with a given level of GFP fluorescence representing specific fractions is shown on the y-axis. Coloured lines with within histogram plots correspond to colours within the summary data shown in (**h**). **h**, Summary of data derived of 8 separate cell sorting experiments that contributed to the data presented in Fig. 3. Graphs show mean+/- SD for n=8 differentiation experiments. **i**, Heat map representation of the expression levels of genes encoding markers used for isolation of specific differentiation stages of differentiation by FACS (see Fig. 3): All negative (AN), RAG1 negative (RN), Double negative (DN), Immature Single Positive (ISP) and Double Positive (DP). The row z-score is an indication of the degree of deviation from the mean for each row. **j**, The relationship between the number of reads mapping to the targeted RAG1 locus containing the GFP gene, the unmodified RAG1 locus and the RAG2 locus for the indicated fractions. RNAseq data corresponds to that analysed in Fig. 3. n=3 or 4, as indicated, samples derived from independent cell sorting experiments (see Source Data for individual values. Because the y-axis is a log scale, samples with no reads are shown as yellow circles below each bar. Note that samples designated as RAG1- generally contained fewer that 10 (10^1) reads. For all flow cytometry plots, except for panels (**d**) and (**f**), numbers on the x and y axis are the exponent of log10 fluorescence. For panel d, the numbers for both x and y axis are the exponent of log10 Biexponential scaling from FlowJo® flow cytometry software, where the (-) indicates -10. The percentage of cells in each quadrant or associated with each gate is indicated. All cells displayed as green are RAG1+. Scale bars for b, from left to right are 200, 600 and 500 μm respectively.



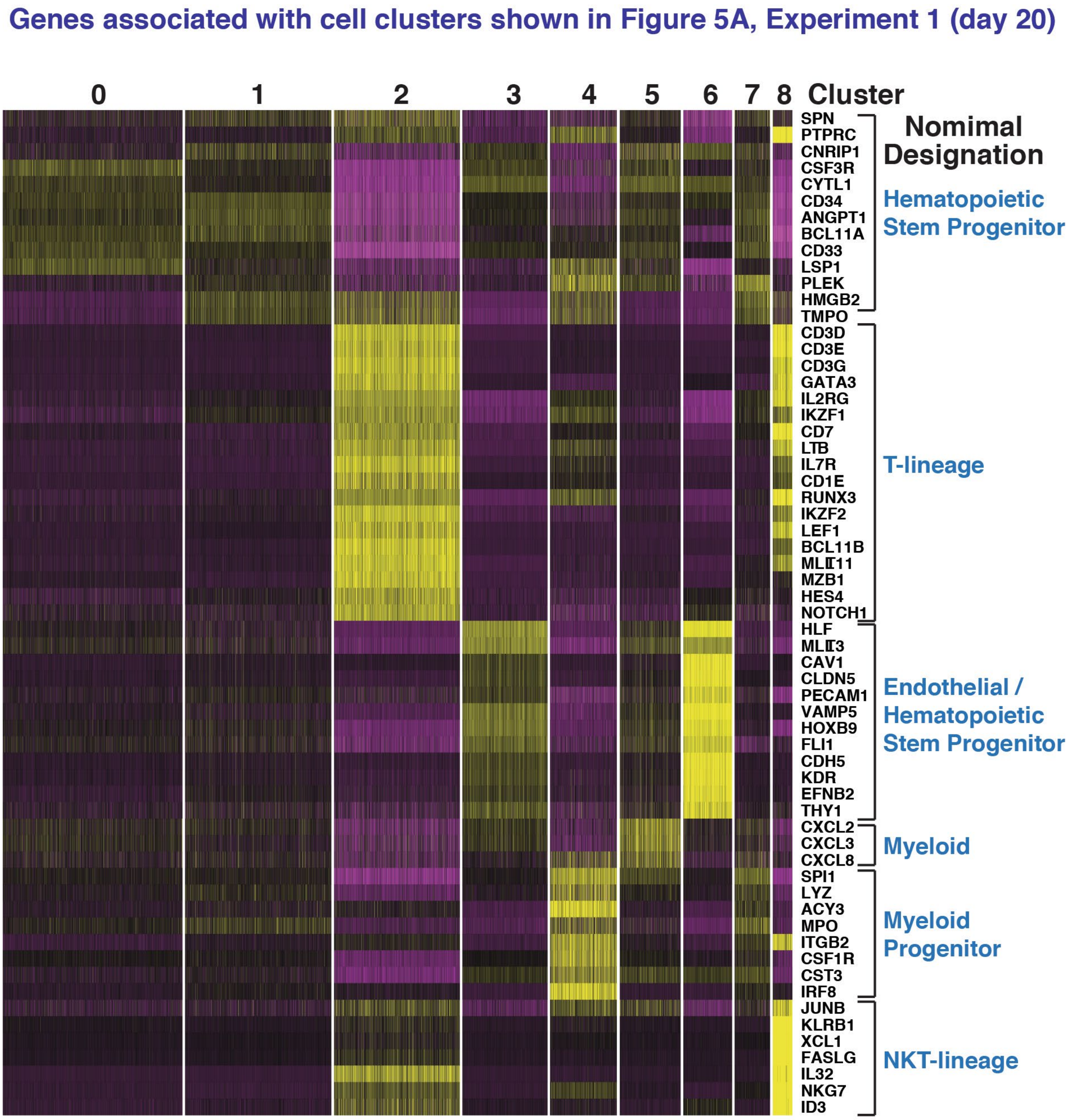
Extended Data Fig. 3 | Characterization of FACS purified subfractions of RAG1+Cellfy. a, Representative gating strategy used for isolation of cell populations profiled in Figs. 4, 5 and Extended Data Fig. 3. **b**, Flow cytometry plots showing the 2 most disparate examples of CD34 and CD90 expression in cells isolated for experiments that form part of the analysis presented in Fig. 4. In these examples the frequency of CD34+CD90+ cells varied from 0 to 10% of the RAG1+ population. **c**, **d**, Relationship between RNA sequencing reads for GFP, RAG1 and RAG2 across cell populations purified on the basis of GFP expression. **c**, Histograms summarizing the number of reads for the indicated genes derived from cell sorting experiments that contributed to data in Fig. 4. For iPSCs, n = 3 or 4, as indicated, For hESCs, n = 1 or as indicated. See source data for details. Because the y-axis is a log scale, samples with no reads are shown underneath each bar as a yellow point. **d**, Trend line representation of the relationship between the of number GFP and RAG1 reads for each sample across all experiments depicted in C above and supplementary figure 2j. n = 38 independent samples derived from cell sorting experiments The equation representing the line of best fit is shown, as is the correlation coefficient for this line ($R^2 = 0.69$). **e**, Heat map representation of mean expression levels of 3′ HOXA genes (A9-A13) across all samples from the indicated fractions for experiments described in Fig. 4.



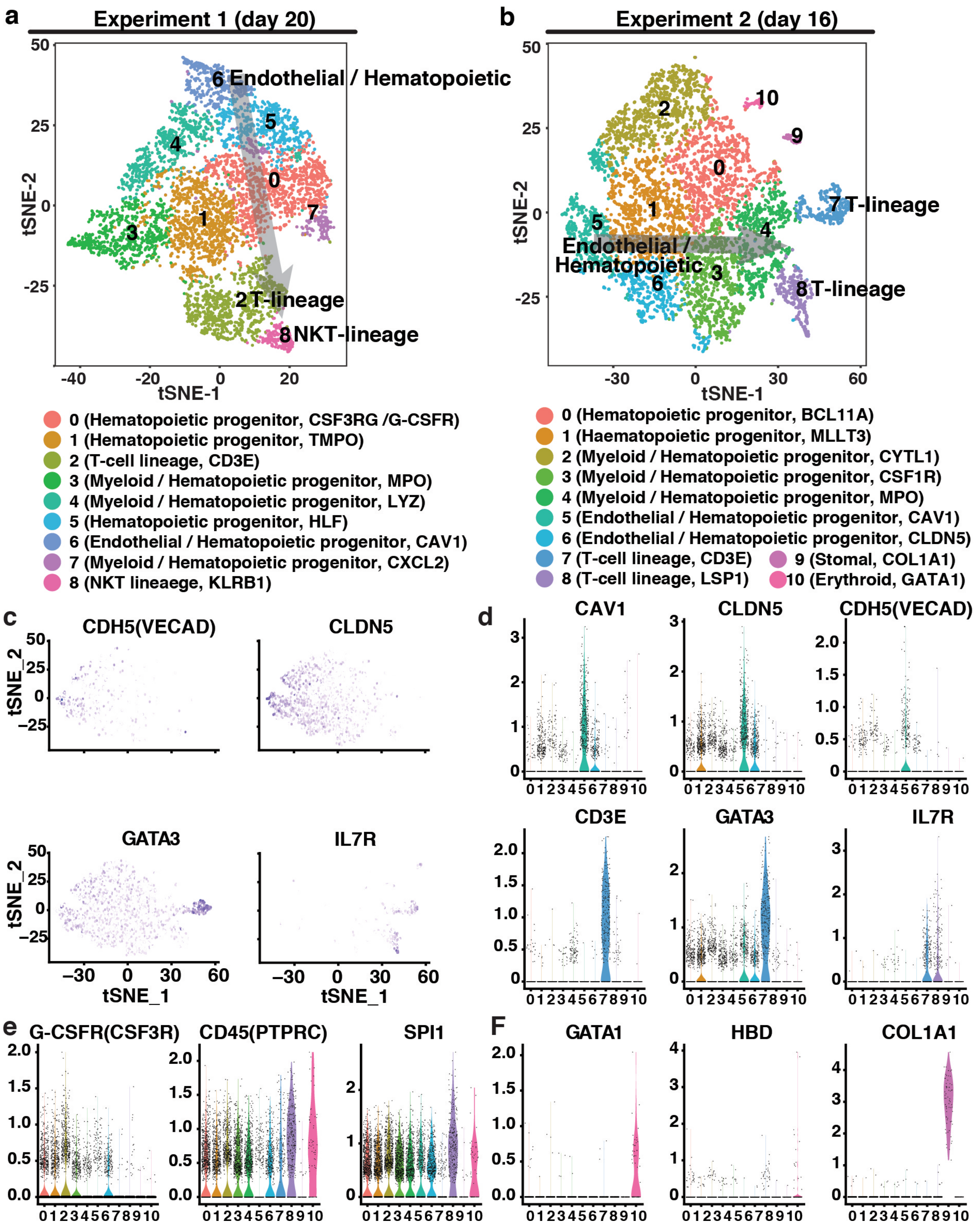
Extended Data Fig. 4 | TCR gene expression of PSC organoid derived T-cell progenitors. a, Graphical summary of the level of gene expression associated with specific TCR loci (see colour key) as indicated. The sorting parameters for individual populations are indicated across the X-axis, and the level of expression for each gene is given as RPKM on the Y-axis. Plots for both the iPSC and hESC based RAG1 reporter cells are shown. The mean expression level for each gene is shown, and where three or more independent samples were collected, error bars represent the SEM. Dots represent data from independent experiments and interconnecting the lines intersect the mean value for the indicated markers for each sorting parameter. For the iPSC and hESC based lines, the number of samples assayed for each fraction were as follows: CD34+CD90+RAG1- (n = 3, n = 1), RAG1+(n = 4, n = 2), CD5 + CD7+CD4-CD8-RAG1+(n = 4, n = 4), CD4+CD8-RAG1+(n = 4, n = 4), CD4+CD8+RAG1+(n = 4, n = 4). **b**, Heat map showing the expression levels of specific TCR genes representing the alpha, beta, gamma and delta loci. Genes included had more than 1 RPKM in 2 different samples. The samples generated using either the iPSC or hESC based reporter lines are indicated with open and filled circles respectively. As indicated by the Row Z-score colour key, genes more highly expressed are shown in red.



Extended Data Fig. 5 | Potential of RAG1+ cells derived from the hESC RAG1:GFP reporter line. a, Methylcellulose colony forming assays for each of the indicated fractions. The number of samples contributing to data for each fraction was RAG1+(n = 3), CD90+CD34+RAG1- (n = 3) and CD90-CD34+RAG1- (n = 2). The mean number of colonies is shown, with error bars representing the s.e.m. The open circles indicate the data values that contributed to the calculation of the mean for each fraction. **b**, Flow cytometry analysis of methylcellulose cultures from the indicated fractions showing the presence of CD235a + (Glycophorin A) erythroid progenitors and CD14+and/or CD16 + myeloid lineage cells. **c**, Flow cytometry analysis of CD45 + blood cells derived from OP9 stromal cultures seeded with organoid derived progenitors for the fractions indicated. Robust numbers of the CD14+myeloid cells are observed in cultures derived from either the RAG1+or CD34+CD90+RAG- fractions. A small number of CD19 + B-lymphoid cells are also observed in these cultures, but not in those seeded with CD34+CD90-RAG1- cells. For all flow cytometry plots using antibody stains, numbers on the x and y axis are the exponent of log10 fluorescence. The percentage of cells in each quadrant or associated with each gate is indicated. All cells displayed as green are RAG1+.

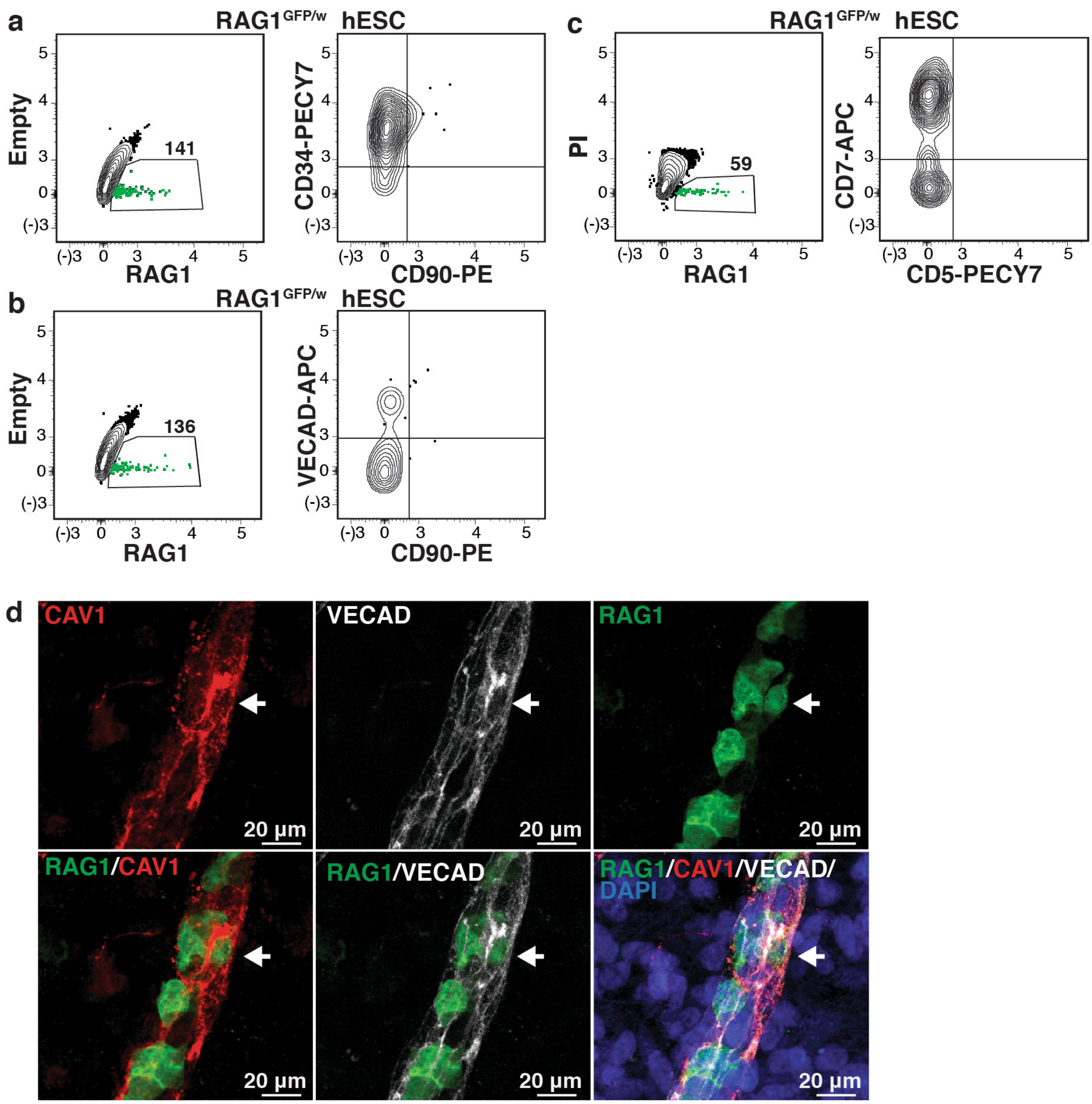


Extended Data Fig. 6 | Single cell sequencing analysis of emergent RAG1+ cells. Heat map showing gene expression levels for key genes in individual cells used in the construction of the tSNE plots shown in Fig. 5a, b, with nominal assignment of lineage associations for each set indicated on the right of the map. The 9 subpopulations (0–8) are indicated across the top of the heat map (Clusters).



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Gene expression analysis of early RAG1+ cells using single cell RNAseq. a, tSNE plot of the data shown in Fig. 5a without prior transformation using MAGIC[42] (van Dijk, D et al 2018), showing that similar populations are present in both forms of the plot. **b**, tSNE plot summarizing the results of a second independent single cell RNAseq experiment (without prior transformation of the data using MAGIC). For both (a) and (b), an assignment of the identity of each group is suggested, along with a single exemplary gene for that cluster which formed part of the set used to assign that cluster. Note that numbering of clusters is based on their relative size, with the cluster designated as "0" containing the largest number of cells. The colouring of clusters was not specified such that similar clusters in each experiment share similar colours. However, the relative position of clusters representing the endothelial/haematopoietic progenitors and T-lineage cells are shared between the two experiments, as indicated by the transparent overlaid transparent grey arrow. **c**, tSNE plots of four genes illustrating the transition of cells from CDH5 + (VE-CAD +) CLDN5 + endothelial/haematopoietic progenitor cells to GATA3 + IL7R + T-cell progenitors in Experiment 2. **d**, Violin plots indicating the relative level of endothelial/haematopoietic progenitor and T-cell progenitor gene expression associated with each cell in each of the fractions as indicated for Experiment 2. The y-axis shows the expression level (lnTPM + 1). **e**, Violin plots displaying the relative levels of expression of the haematopoietic progenitor and myeloid marker G-CSFR and the myeloid marker SPI1 across the CD45 + (PTPRC) blood-cell fractions. **f**, Violin plots showing the expression of the stromal marker (COL1A1) and erythroid genes (GATA1 and HBD) in the minor contaminating populations, 9 and 10. For d, e and f, dots represent individual data points (single cells), with the position on the y-axis representing the relative expression of the indicated gene in that cell. N numbers in d-f represent the individual cells (data points) with a given expression value for the indicated gene and are shown as individual data points in the plots.



Extended Data Fig. 8 | Flow cytometry and immunofluorescence of early RAG1+ cells. a, **b**, Flow cytometry plots showing that a proportion of emergent hESC-derived RAG1+ cells co-express CD90 and CD34 (a) but, in contrast to iPSC RAG1+ cells, only a small number of VE-CAD + cells retain CD90 expression (b). **c**, Flow cytometry plot showing a proportion of early RAG1+ cells are CD5-CD7-. For a, b and c, the number (a:136, b:141, c:59) and position of cells falling within the RAG1+gate that were used to generate accompanying contour plots is indicated. The scale numbers for both x and y axis are the exponent of log10 Biexponential scaling from FlowJo® flow cytometry software, where the (-) indicates -10. **d**, **e**, Confocal immunofluorescence analysis showing the expression of RAG1, VE-CAD and CAV1 (**d**), and RAG1 and GATA3 (**e**) on cells localized or confined to defined structures. The presence of rare RAG1+ cells that co-express CAV1 and VE-CAD (marked by the arrow in **d**) is consistent with the initiation of RAG1 expression within cells that form part of the vascular structure. Nuclei were identified by staining with DAPI. Scale bars = 20 μm.

nature research

Corresponding author(s): Ed Stanley

Last updated by author(s): Nov 22, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Software used in this study includes ZEN blue (Zeiss, version 2.1), ImageJ (Version 1.51) and Imaris software (Version 9.2.1).
Data analysis	Images were exported as JPEG or PNG files and labelled using Adobe illustrator (Version 23.1). Analysis of data sets from flow cytometry and selected RNAseq experiments was performed using Microsoft Excel (version 16.30) of Prism8 (Version 8.2.1 (279). Single cell sequencing data was analyzed with Cell Ranger (Version 3.0.2). Details of the source of non-commercially available software used to analyze RNA sequence data is detailed and referenced in the Methods (supplementary materials)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All single cell RNA-seq and bulk RNA-seq data used in this study have been deposited in the Gene Expression Omnibus (GEO) database repository, and are available under GEO accession numbers GSE124172 (scRNA-seq) and GSE124173 (bulk RNA-seq). All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Independent experimental replicates were used to validate specific findings where this was deemed necessary. In general, sufficient experimental replicates were performed to enable a mean and standard deviation or SEM to be calculated. However, single experiments were also performed to confirm that specific results could be reproduced using a separate cell line. In all cases the number of experiments and is provided in the figure legends and/or supporting information files.
Data exclusions	No Data was excluded from this study
Replication	Independent experiments were used replicate key findings in this manuscript. In addition, 2 independently derived reporter lines were used as a second level of experimental validation. The number replicates for data presented in each figure are indicated in the relevant legend or the accompanying text.
Randomization	This study did not involve sub-sampling a large population and therefore randomization was not relevant.
Blinding	Investigators were not blinded for data collection or analysis because the intent of both collection and analysis was not to test a null hypothesis

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☒	☐	Human research participants
☒	☐	Clinical data

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used	All antibodies used in experiments in this study can be found in Supplementary Table 2.
Validation	Validation of the large number of antibodies used was based on a combination of the manufacturers' validation statements accompanying the antibodies and our own experience in using the antibodies.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The RM3.5 iPSC line used in this study was generated by our laboratory and published in Kao et al, 2016 (Kao, T. et al. GAPTrap: A Simple Expression System for Pluripotent Stem Cells and Their Derivatives. Stem Cell Reports 7, 518-526 (2016). The H9 human embryonic stem cell line used in this study was obtained from WiCell and published by Thomson, J.A. et al. Embryonic stem cell lines derived from human blastocysts. Science 282, 1145-1147 (1998).
Authentication	The authenticity of the cell lines were checked by SNP analysis following genetic manipulations to ensure they possessed a normal karyotype that was consistent with that of the parental line.

Mycoplasma contamination

Pluripotent stem cell lines were routinely tested and found to be negative for mycoplasma by the Cerberus Sciences, Adelaide,

Commonly misidentified lines
(See ICLAC register)

Commonly misidentified lines were not used in this study

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Cells were harvested from pluripotent stem cell differentiation cultures using enzymatic disaggregation to produce a single cell suspension. Following harvesting and antibody labeling, cell suspensions were passed through filter cap of flow cytometry tubes to eliminate large clumps. Cells were collected by centrifugation and washed to remove unbound antibodies. After a final centrifugation step, cells pellets were resuspended in FACS buffer containing propidium iodide.

Instrument

Flow cytometric analysis was performed using a BD LSR Fortessa analyser. Flow sorting used a BD Biosciences InfluxTM or BD Biosciences FACSAriaTM Fusion cell sorter.

Software

Flow Flow cytometry data was collected and analysed with BD Diva software (Version 8.0.1). Processing and preparation of data for presentation in figures was performed using either Flowlogic (Version 2.2) or Flowjo software (Version 10.5.0).

Cell population abundance

For flow cytometry analysis, the percentage of cells present in each fraction or quadrant is indicated, if relevant to the interpretation of the plot.

Gating strategy

For flow cytometry analysis, gates were originally set using either an isotype control antibody or using samples labeled with a set antibodies lacking a fluorochrome on for the channel for which the gate was to be set. Gates for GFP+ cells were set against cell types known to be negative for GFP expression or cells derived from PSC lines not harboring the specific GFP reporter gene.

For flow cytometry sorting, (bulk RNA sequencing analysis), an example of the gates, as used to define specific fractions, is presented in figures 3A and 4A. Note that fractions described in figure 4A were also for methyl-cellulose cultures shown in figure 5. The purity of the sort fractions was generally >90%, as determined by re-analysis of a small amount of the sorted sample. This was not tested for every fraction in every experiment, particularly in instances where cell numbers were low or limiting.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.