Functional and transcriptional heterogeneity of human hemopoietic lympho-myeloid progenitors at the single cell level

Dimitris Karamitros1,2¶, Bilyana Stoilova1,2¶, Zahra Aboukhalil1, Fiona Hamey3§, Andreas Reinisch4§, Marina Samitsch1, Lynn Quek1,2,5, Georg Otto1, Emmanouella Repapi1, Jessica Doondeeda1,2, Batchimeg Usukhbayar1,2, Julien Calvo6, Stephen Taylor1, Nicolas Goardon1, Emmanuelle Six7, Francoise Pflumio6, Catherine Porcher1, Ravindra Majeti4, Berthold Gottgens3, Paresh Vyas1,2,5*

1MRC Molecular Haematology Unit and 2Oxford Biomedical Research Centre, Weatherall Institute of Molecular Medicine, University of Oxford, UK; 3Department of Haematology and Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK; 4Division of Hematology, Stanford Institute for Stem Cell Biology and Regenerative Medicine, Stanford, CA, USA; 5Department of Hematology, OUH NHS Trust, UK; 6UMR967 INSERM/CEA/Université Paris 7/Université Paris 11, Paris, France; 7UMR1163, Paris Descartes–Sorbonne Paris Cité University, Imagine Institute, Paris, France. ¶These authors contributed equally to this work and are listed alphabetically. §These authors contributed equally to this work.

*Corresponding author: Paresh Vyas, MRC Molecular Haematology Unit and Department of Haematology, Weatherall Institute of Molecular Medicine, University of Oxford OX3 9DS. Telephone: +44-1865-222309. Fax: +44-1865 222501. E-mail: paresh.vyas@imm.ox.ac.uk.

Abstract 117 words

Total length: 6089 words
Abstract: The human hemopoietic progenitor hierarchy producing lymphoid and granulocytic-monocytic (myeloid) lineages is unclear. Multiple progenitor populations produce lymphoid and myeloid cells, but they remain incompletely characterized. Here, we demonstrate current cord blood lympho-myeloid containing progenitor populations - the lymphoid-primed multi-potential progenitor (LMPP), granulocyte-macrophage progenitor (GMP) and multi-lymphoid progenitor (MLP) - are functionally and transcriptionally distinct and heterogeneous at the clonal level, with progenitors of many different functional potentials present. Though most progenitors have uni-lineage myeloid or lymphoid potential, bi- and rarer multi-lineage progenitors occur in LMPP, GMP and MLP. This coupled with single cell expression analyses, suggests a continuum of progenitors execute lymphoid and myeloid differentiation rather than only uni-lineage progenitors being present downstream of stem cells.

Human hemopoiesis produces 10 billion new, terminally mature, blood cells daily; a production that is also rapidly responsive to external change. Most of this production generates red cells required for oxygenation, shorter-lived myeloid cells (granulocytes, monocytes and dendritic cells) and platelets. It also replenishes long-lived acquired immune cells (T and B lymphocytes) and innate immune natural killer cells. Dysregulation of this complex process can lead to hemopoietic and immune deficiencies and blood cancers.

A complex cellular hierarchy sustains this enormous cellular production. Despite being one of the best-studied tissue hierarchies, active debate continues about the heterogeneity and plasticity of hemopoietic cell populations, in steady state and in response to stimuli. At the hierarchy apex lie multi-potent hemopoietic stem cell (HSC) populations (more fully characterised in mice than human) heterogeneous with respect to differentiation potential, cell cycle, self-renewal capacity, stability over time and contribution to hemopoiesis in steady state versus transplantation. Downstream of murine long-term HSCs are heterogeneous short-term HSC (HSC<sup>ST</sup>), multipotent (MPP) and early lineage-biased progenitors. In human the HSC<sup>ST</sup>/MPP population has not been fully defined. In
terms of lineage potential restriction, there is increasing support that the erythroid and megakaryocyte fates diverge from other myeloid and lymphoid potentials in mouse\textsuperscript{14, 17, 18, 19, 20} and man\textsuperscript{21, 22, 23, 24, 25}. Commensurately, erythroid and megakaryocytic cells may differentiate directly from either HSC\textsuperscript{6} or immediate downstream MPP\textsuperscript{14, 16, 26}.

Focusing on the first human lympho-myeloid progenitors downstream of HSC and MPP, two progenitor populations have been identified within the immature Lin\textsuperscript{−}CD34\textsuperscript{+}CD38\textsuperscript{−}CD90\textsuperscript{−/lo} compartment (Fig1ai). These include a Lin\textsuperscript{−}CD34\textsuperscript{+}CD38\textsuperscript{−}CD90\textsuperscript{−/lo}CD45RA\textsuperscript{+}CD10\textsuperscript{−} lymphoid-primed multi-potential progenitor (LMPP) with granulocytic, monocytic, B- and T-cell potential, but unable to generate erythroid or megakaryocytic cells\textsuperscript{22}. These data support previous studies showing human CD34\textsuperscript{+}CD10\textsuperscript{−} cells retain lympho-myeloid potential but progressively lose myeloid potential with CD10 expression\textsuperscript{27, 28}. In contrast, the multi-lymphoid progenitor (MLP) is CD10\textsuperscript{+} with lymphoid (B, T, NK), monocytic and dendritic cell potential but cannot make granulocytes\textsuperscript{21}. Within the Lin\textsuperscript{−}CD34\textsuperscript{+}CD38\textsuperscript{−}CD45RA\textsuperscript{+} compartment, there are again at least two lympho-myeloid progenitors. One is a CD62L\textsuperscript{hi}CD10\textsuperscript{−} lymphoid-primed progenitor with lymphoid, monocytic and dendritic cell potential\textsuperscript{23} and the granulocyte-monocyte progenitor (GMP; Lin\textsuperscript{−}CD34\textsuperscript{+}CD38\textsuperscript{−}CD45RA\textsuperscript{+}CD123\textsuperscript{+}) which has mainly myeloid potential but retains residual lymphoid potential\textsuperscript{22, 29} consistent with the residual lymphoid potential in the murine pre-GM progenitor\textsuperscript{30}. Finally, the human Lin\textsuperscript{−}CD34\textsuperscript{+}CD38\textsuperscript{−}CD45RA\textsuperscript{+} compartment also contains a CD10\textsuperscript{+} subpopulation with T-, B-, NK and dendritic cell potential but lacking other myeloid cell potential\textsuperscript{31}, consistent with the notion that acquisition of CD10 expression accompanies loss of some, or all, myeloid potential.

These prior observations raise questions about whether these progenitor populations are pure or heterogeneous, how distinct they are and the nature of the functional, transcriptional and hierarchical relationships between them (Fig1ai-iii). Without answers to these questions, our understanding of human hemopoiesis will remain severely limited. Here, we directly and rigorously compare the in vitro and in vivo functional potential and transcriptional programs of the human LMPP, MLP and GMP at a population and single cell level. Using single cell
functional assays we show that the majority of progenitors in these populations have a variety of uni-lineage potentials. However, a substantial minority of progenitors retains diverse bi-lineage potentials, while 6-15% of both the LMPPs and GMPs have combined lympho-myeloid potential at a single cell level. Thus, multi-lineage lympho-myeloid potential is retained quite late in the hierarchy of currently purified progenitor populations.

Results

Immunophenotypic purification of lympho-myeloid progenitors

We improved prior FACS-staining and sorting strategies to prospectively purify eight human hemopoietic stem/progenitor cell (HSPC) populations, including the LMPP, GMP and MLP. \(^{22}\) (Fig. 1b-c, Supplementary Fig. 1a-b) in human cord blood (CB) and bone marrow (BM). In our gating strategy, we included all Lin$^-\text{CD34}^+$ cells; the eight HSPC populations accounted for 82% of all CB Lin$^-\text{CD34}^+$ cells (Figure 1). The remaining cells did not constitute separate populations. Critically, the Lin$^-\text{CD34}^+\text{CD38}^-\text{CD45RA}^+$ compartment contains a mixture of CD10$^-\text{LMPP}$ and CD10$^+\text{MLP}$ progenitor populations. Furthermore, the more mature Lin$^-\text{CD34}^+\text{CD38}^+$ compartment is separated into the CD10$^+\text{B-NK}$ containing progenitor population\(^{31}\) and CD10$^-$ heterogeneous myeloid progenitors (Fig. 1c-e).

Immunophenotypic LMPP and MLP are very rare (Fig. 1d-e). Using analysis gates they each constitute 0.2% of the BM Lin$^-\text{CD34}^+$ compartment. As Lin$^-\text{CD34}^+$ comprises ~1% of mononuclear cells (MNCs), LMPP and MLP are present at ~2/10^5 BM MNCs. Though more frequent in CB (LMPP 1.4% and MLP 1.1% of Lin$^-\text{CD34}^+$ cells), they still constitute only ~1/10^4 MNCs in CB. GMPs are 20-fold more abundant in CB and 100-fold more abundant in BM than LMPPs and MLPs, comprising 16.8% and 17.8% of Lin$^-\text{CD34}^+$ cells in CB and BM respectively (i.e. ~1.5-2/10^3 MNCs).

Human lympho-myeloid populations have distinct functional potential in vitro

As the frequency of adult BM LMPP and MLP was extremely low, all functional assays performed hereafter were with fresh CB cells. HSPC populations (HSC, MPP, LMPP, MLP,
CMP, GMP, MEP) were double sorted to high purity (>99% except the CMP (97%), Supplementary Fig. 1b). First, we performed methylcellulose-based colony forming unit (CFU) assays to quantitatively compare clonogenic myeloid potential (Fig. 2a-b). In comparison to GMP (31% cloning efficiency), LMPP and MLP had lower myeloid clonogenic potential (6% and <1%, respectively) consistent with previous published data15, 21, 22. GMP and LMPP generated granulocytic (G), monocyte/macrophage (M) and GM colonies with either no/or trace erythroid (E) potential (<0.5%) as previously described22, 32. MLP generated very few myeloid colonies (only monocyte). As control, E, M, G, GM and mixed colonies were detected from CMP and MPP populations16, 22, 32.

Next, we analyzed lymphoid and myeloid differentiation potential of 150 LMPP, MLP and GMP cells with an in vitro liquid differentiation assay on MS-5 stroma. We initially performed a kinetic analysis of different lineage outputs to determine the optimal time point to detect hCD45+CD15+ neutrophil (G), hCD45+CD14+ monocyte (M), hCD45+CD19+ B-cell and hCD45+CD56+ NK cell differentiation (Fig. 2c and Supplementary Fig 2a). From LMPP, MLP and GMP, G production was detected at weeks 1 and 2 but was maximal at week 1. Contrastingly, NK and B lymphoid production was present at weeks 2 and 3 but maximal at week 3, when granulocyte production had ebbed. Monocyte production was detected maximally at weeks 1 (GMP) or 2 (LMPP and MLP). We confirmed the lineage identity of mature FACS purified G, M, B or NK cells by testing expression of granulocytic (CSF3R, MPO, TREM1, CEBPA), monocytic (SPI1, CSF1R, CX3CR1), B-cell (EBF1, CD79A, POU2AF1, PAX5) and NK cell (NCAM1, CD2, CD3E, GATA3) lineage affiliated genes (Fig. 2d). Given these kinetics, we analyzed subsequent limiting dilution and single cell cultures at week 2 to capture all four terminal myeloid (G and M) and lymphoid (B and NK) outputs.

Next, we tested if LMPP, GMP and MLP populations generated T cells in vitro at weeks 5 and 7 on OP9-hDL1 stroma (culture condition SF7a) (Fig 2e, Supplementary Fig. 2b). All 3 progenitors generated both immature (hCD7+CD1a+) and more mature (DP: hCD7+CD1a+hCD4+CD8+; ISP: hCD7+CD1a+CD4+CD8+, hCD7+CD1a+CD4+CD8+) T-cells at
both time points but with greater production at week 5. FACS purified T-cell subpopulations expressed CD2, CD3E, GATA3, LEF1, RAG1, NOTCH1 and IKZF1, confirming their lineage affiliation (Fig 2f).

**Single cell clonal analyses reveal functional differences and heterogeneity of LMPP, MLP and GMP**

To quantitatively test clonal functional potential we used three, different optimized in vitro liquid culture assays to exhaustively test lympho-myeloid potential (**Supplementary Figure 3a**). First, we established frequencies with which LMPP, MLP and GMP produce lymphoid (B and NK) and myeloid cells (M and G) by limit dilution assay (LDA) (with 1, 2, 5, 10 and 20 cells) in a new efficient liquid culture assay on MS-5 stroma with cytokines (SGF15/2 condition). Lineage output was assessed by FACS (**Fig. 3a-b, Supplementary Figure 3b-c, Supplementary Table 3a**). One in 2 LMPP cells produced B-cells, 1 in 3 differentiated into NK cells but only 1 in 5 LMPP cells differentiated into monocytes and 1 in 10 into granulocytes. GMPs generated myeloid cells with higher frequency (1 in 2 differentiated into monocytes and 1 in 4 into granulocytes). GMPs also produced lymphoid cells at a much lower frequency (NK cells – 1 in 8 and B-cells – 1 in 22). MLP were also lymphoid-biased (1 in 11 MLP cells produced B-cells and 1 in 18 generated NK cells) whereas myeloid output was rare (1 in 194 differentiated into monocytes and 1 in 394 generated granulocytes). Finally, bi-lineage and multi-lineage cells were generated with much lower frequencies.

Thus, the three populations are functionally different. However, the LDA does not rigorously define frequency of multi-lineage functional potential at clonal level. Therefore, we studied in vitro lympho-myeloid (G, M B, and NK) potential of 3468 single cells (1136 LMPPs, 710 MLP and 1622 GMPs cells, **Supplementary Table 3b**) from 22 biological CB replicates representing a total of 6.3x10⁹ MNCs providing robust quantitative data especially for rare functional potentials within the infrequent lympho-myeloid progenitor populations.

At a single cell level LMPP and GMP had high cloning efficiency (54% and 71% respectively), whereas clonogenicity of MLP was 11% (**Fig 3c**). The LMPP and MLP were
primarily lymphoid progenitors and GMP mainly a myeloid progenitor, consistent with the LDA data (Fig. 3di-ii). Focusing on wells that gave single cell initiated cultures 69% of LMPP, 88% of MLP and 63% of GMP gave uni-lineage output (Fig. 3e). When there was uni-lineage output, 92% of LMPP cells had lymphoid output (B or NK cells) and 8% myeloid output (G or M). The MLP was virtually exclusively a lymphoid progenitor with very low myeloid output (3%). In contrast, 79% of GMP cells had myeloid output and 21% lymphoid output. Bi-lineage output was the next most common potential (24% of LMPP, 12% of MLP and 33% of GMP) (Fig. 3f). Output of three or more lineages was rare (6% of LMPP, 0% of MLP and 3% of GMP) (Fig. 3g). Regardless of whether output was uni-, bi- or multi-lineage, 15 different lineage-affiliated progenitor types were detected. However, only 8% of all plated LMPPs, 7% of GMPs and hardly any MLP (0.3%) exhibited combined lympho-myeloid potential (Fig. 3h). Fisher's exact test demonstrated that the lympho-myeloid output from LMPP was significantly higher compared to GMP (p=0.0125) and MLP (p=0.0019, Supplementary Table 3c).

We also tested 258 single cells with alternative culture (Condition 2 (S7T2GM/G/M) Supplementary Fig. 3a) used to define the MLP\textsuperscript{21}. The results were similar to those with the first culture condition but were much less permissive, especially for granulocyte output. Most output was uni-lineage with rarer bi-lineage and multi-lineage outputs (Supplementary Figure 3d-i, Supplementary Table 3d). The LMPP contained most of the lympho-myeloid potential; the MLP was a lymphoid-biased progenitor and the GMP a myeloid-biased progenitor.

Finally, we studied 631 single LMPP, MLP and GMP with a third culture condition (SF7b/Dox) optimized for combined T-lympho-myeloid cell potential on MS-5/hDL-1\textsuperscript{IND} stroma where hDL1 expression was induced after 3 weeks of culture (Fig 3i-n, Supplementary Table 3e) by adding doxycycline to the culture. We analyzed B/NK/M/G output after 3 weeks and T-cell output after another 3 weeks. Again, a large diversity of progenitors was detected; uni-lineage T-cell output was detected in LMPP and MLP
populations (3% of positive wells for both populations) and was virtually absent in GMP (<0.1%) (Fig 3k). T-cell combined with other lymphoid output (NK, B-cells, or both) was detected in a minority of cells (1-5%) in LMPP, MLP and rarely in the GMP (Fig 3k). Most strikingly, lympho-myeloid output was only seen in the LMPP (14%) (Fig 3m). Overall, 24 different progenitor types were detected by the in vitro analyses; all of which were seen in the LMPP and subsets were detected in MLP and GMP (Fig 3o).

**Human LMPP, MLP and GMP progenitors have distinct differentiation potential in vivo**

We assayed the functional potentials of these short-lived progenitor populations in immunodeficient mice, acknowledging that in vivo single cell clonal assays of human progenitors are not yet feasible. Previous xenotransplantation assays have yielded very low (<0.1%) or no engraftment of MLP and LMPP\(^\text{15, 21, 22, 25}\). We used a novel in vivo functional assay, where progenitors were transplanted into humanized ossicles, in immunodeficient mice\(^\text{33}\) (Fig. 4a and Supplementary Fig. 4a-b). To better understand in vivo kinetics of LMPP, MLP and GMP engraftment, we initially analyzed ossicles 1 and 2 weeks after transplantation. Engraftment was detected at both time points with greater hCD45\(^+\)hCD33\(^+\)hCD14\(^+\) (M), hCD45\(^+\)hCD33\(^+\)hCD15\(^+\) (G), hCD45\(^+\)hCD33\(^-\)hCD19\(^+\) (B) engraftment at week 2 (data not shown). Therefore, subsequent analyses are reported 2 weeks post injection.

Direct transplantation into humanized ossicles allowed \(~10\text{-}100\)-fold more engraftment of human progenitors (Fig 4b) than previously reported\(^\text{21, 23, 25}\). As the number of injected cells varied per population (mean 9618 LMPP cells, 1588 MLP cells and 20500 GMP cells) (Supplementary Fig. 4c), we report mean human cell engraftment per 1000 transplanted cells. The GMP population had the highest human cell engraftment (2.6±0.88%; mean±1SEM) followed by LMPP (1.4±0.5%) and then MLP (0.2±0.06%). Transplanted LMPP produced more CD33\(^+\) cells (mean 82±3.3%) than B-cells cells (17±3%) (Fig 4c-d). In contrast, transplanted MLP generated more B-cells (78±5.9%) than CD33\(^+\) cells (19±6.7%). There was no correlation between the number of transplanted LMPP and MLP cells and the
lympho-myeloid output ratio (Supplementary Fig. 4d). The majority of differentiated progeny of GMP were CD33+ cells (97±0.7%, Fig. 4c-d). When further analyzed, engrafted CD33+ cells from the LMPP and GMP expressed monocytic (CD14) and granulocytic (CD15) markers; but no CD14 and/or CD15 expressing cells were detected from the MLP (Fig. 4d). Giemsa staining confirmed that CD15+ cells were granulocytic and CD14+ monocytic (Fig. 4e). CD14+CD15+ were more immature myeloid cells. Strikingly in vivo only the LMPP has lympho-myeloid potential but with more myeloid than lymphoid output. In contrast, the GMP has principally myeloid potential and the MLP limited lymphoid potential. The in vivo functional assays reinforce in vitro observations that the LMPP, MLP and GMP are functionally different.

**Distinct transcriptional patterns of lympho-myeloid progenitors correlate with function**

We sequenced RNA from 3-4 independent replicates of the human HSPC populations (HSC, MPP, LMPP, MLP, CMP, GMP and MEP), asking the following questions: i) How do the LMPP, MLP and GMP populations transcriptionally relate to each other and to other HSPC populations? ii) Can we define a distinct gene expression signature for each lympho-myeloid progenitor population? iii) Can the gene expression profile of the three progenitors be correlated with their functional output (i.e. do progenitors exhibit lineage priming)?

Hierarchical clustering with all 14646 expressed genes separated the LMPP and MLP from other HSPCs and the HSC/MPP clustered away from mature progenitors (Fig. 5a and Supplementary Fig. 5a). To define the transcriptional similarities and differences between HSPC populations we used ANOVA analysis to obtain differentially expressed genes (DEG) (Supplementary Table 4). We then performed Principal Component Analysis (PCA), using all expressed genes or between 300 and 10000 of the most DEG (ranked by p-value, adjusted for multiple testing) (Fig. 5b). By focusing on DEG the separation between populations became clearer. Using DEG, the HSPC populations separate along principal component (PC) 1 by lineage potential (lymphoid on the left and erythroid/megakaryocytic on the right with GM in the middle). PC2 separates populations by maturation. The best
separation between these functionally distinct populations was seen with the 300 most DEG. Next, we determined the eigenvalues of the principal components using these 300 most DEG versus a random sampling of 300 genes from a randomized data set. This showed that the first three PCs captured most of the transcriptional variation between the populations (Supplementary Fig. 5b-c).

As ANOVA analysis presumes that HSPC populations are distinct we reanalyzed our RNA-seq data identifying genes with the highest variance across all the biological populations and replicates without assuming population identity. Sampling 10000, 3000, 1000 and 300 of the most variant genes, we used PCA to display the relatedness of the samples (Fig. 5b and Supplementary Fig 5d). These PCAs were very similar to those generated using the ANOVA gene sets. The top 300 variant genes produced a PCA plot that was virtually superimposable to that generated by the top 300 DEG identified by ANOVA.

The loadings plot for the PCA using the 300 ANOVA gene set identified stem-affiliated genes (HLF, MECOM, NFIB), lymphoid-affiliated genes (IGJ, IRF8, MME) and erythroid/megakaryocytic-affiliated genes (HBD, HPGDS) (Fig. 5c). We also visualized the genes driving transcriptional variation across the populations within the 300 ANOVA DEG by hierarchical clustering (Fig. 5d). Focusing on the LMPP, GMP and MLP, ELANE, MPO and PRTN3 are strongly expressed only in the GMP, whereas the LMPP and MLP share expression of many lymphoid affiliated genes including IL7R, LCK, SYK, ADA, HLX, LST1 and ITGAL.

The transcriptional similarities and differences between HSPC populations were further analyzed through pairwise comparisons (Fig. 5e, Supplementary Fig. 5e, Supplementary Table 5a-g). The number of differentially expressed genes between populations is one measure of the transcriptional distance between HSPC populations, without assuming any hierarchical relationships. The most closely related populations are the HSC and MPP, with only 13 DEG between them. The LMPP and MLP are closely related at a transcriptional level (85 DEG), as are the CMP and MEP (63 DEG). The GMP is most closely related to the CMP.
Given the functional differences between LMPP, MLP and GMP we reasoned that distinct gene expression signatures could be derived for the three populations (Fig. 5f). We used the DEG in one versus all population comparisons and then filtered for unique genes to obtain gene signatures (Fig. 5f i-iii, Supplementary Table 6a-c). The GMP gene signature contained many myeloid genes also identified by ANOVA (Fig. 5d), whereas the MLP signature contained many lymphoid genes (e.g. TCF4, TCF7, RUNX3, IL7R, LCK). By contrast, the gene expression signature of LMPP contained both lymphoid and myeloid genes (lymphoid: ETS1, EBF1, CYTIP; myeloid: TRPM2, S100A8, PADI4, ALOX15B).

To validate these findings, we studied the expression of recently published gene sets affiliated with different blood lineages to score lineage bias and strength of commitment of LMPP, GMP and MLP (Supplementary Fig. 5f). In the GMP genes expressed in immature myeloid, neutrophil and monocyte-dendritic cell lineages were prominent. In contrast, LMPP and MLP expressed genes affiliated with B-cells, monocyte-dendritic cells and not neutrophils (with the exception of FOSB). Finally, we examined enrichment of MetaCore Pathway Maps as another means to test how global gene expression relates to the function (Supplementary Table 7). The GMP was enriched for pathways associated with myeloid maturation (e.g. regulation of granulocyte development: FDR=0.0136). In contrast, the MLP was enriched for pathways associated with lymphopoiesis (e.g. Notch signaling pathway: FDR<0.001). The LMPP had more balanced enrichment for both lymphoid and myeloid associated pathways (e.g. M-CSF signaling pathway: FDR<0.001 and BCR signaling pathway: FDR=0.049).

Given that the global transcriptional programs of LMPP, MLP and GMP mirror their functional potentials, we used two approaches to identify transcription factors (TFs) driving the distinct transcriptional programs. First, we interrogated our lists of genes differentially expressed between the MLP and GMP for differentially expressed TFs (Supplementary Fig. 5g). Second, we validated these observations by examining expression of previously identified hematopoietic TFs in LMPP, MLP and GMP (Fig. 5g).
expressed mainly myeloid TFs (e.g. ERG, GATA2, MYB, EGR1), while lymphoid TFs (e.g. HES1, RUNX3, POU2F2, LEF1, IKZF1, IRF8, TCF4) showed highest expression in MLP. Mirroring its functional abilities, the LMPP showed more balanced expression of both myeloid and lymphoid TFs. A similar trend was seen with expression of cytokine and chemokine receptor associated genes (Supplementary Fig. 5h). In summary, transcriptomic data confirm that LMPP, MLP and GMP are transcriptionally distinct from each other and the other HSPC populations. They have TF and lineage-affiliated signatures consistent with their functional potential.

Single cell gene expression reveals a continuum of lympho-myeloid progenitors

To begin to separate transcriptionally and functionally distinct progenitors within the heterogeneous GMP, LMPP and MLP populations we index FACS sorted single cells for: (i) functional analysis; (ii) RNA-Seq and (iii) quantitative RT-PCR (qRT-PCR). We then used the index data to correlate function and transcriptional state (Fig. 6a).

We first profiled the expression of 96 genes (Supplementary Table 8) in 919 single LMPP, MLP and GMP cells. These genes encoded lineage-affiliated transcriptional regulators and cell surface and lineage-affiliated markers. Genes with low variance and levels of detection were excluded. Thus, expression of 74 genes was taken forward for further analysis.

Hierarchical clustering of transcriptional profiles assigned GMPs, LMPPs and MLPs to three clusters (Supplementary Fig. 6a-bi). Cluster 1 (543 cells) was mainly composed of MLP and LMPP. Cluster 2 (150 cells) was a mixture of GMPs, LMPPs and MLPs. Cluster 3 (226 cells) was mainly composed of GMPs. Transcriptionally, cluster 1 showed higher expression of lymphoid-affiliated genes (e.g. TCF7, EBF1, NOTCH3, CD22, CD79A, MME (CD10) CD3E and CD6) compared to the other two clusters (Supplementary Fig. 6bi). Conversely, cluster 3 showed increased expression of myeloid genes (CEBPA, GFI1, MYB, ID2, CSFR1, CSFR3, MPO and PRTN3, Supplementary Fig. 6bi). Thus, cluster 1 contains more lymphoid-biased cells; cluster 2 a mixture of lympho-myeloid cells; and cluster 3 contains myeloid-biased cells. Finally, functional output of the progenitors in each cluster mirrored the transcriptional
programs seen at a single cell level (Supplementary Fig. 6biii). Thus, cluster 1, a mix of LMPP and MLP (Supplementary Fig. 6bi), produced mainly a lymphoid output (Supplementary Fig. 6biii); whereas cluster 3 composed mainly of GMP produced mainly a myeloid output (Supplementary Fig. 6biii).

Next, we performed dimensionality reduction on the gene expression data using the diffusion map method adapted for single-cell data\(^\text{37}\) we had previously applied\(^\text{38}\). By retrospectively indicating progenitor identity on the diffusion map (Supplementary Fig. 6ci), the MLP, LMPP and GMP cells can be seen to form a continuum (MLP left, LMPP in the middle and GMP right) in agreement with the hierarchical clustering in Supplementary Fig. 6a. We next colored the diffusion map by cluster assignment (Supplementary Fig. 6cii). This demonstrated that cluster 2 lies between clusters 1 and 3, in agreement with its mixed lympho-myeloid transcriptional signature (Supplementary Fig. 6bii).

To guard against potential limitations of gene selection bias from the qRT-PCR data, we performed single cell RNA sequencing of human hemopoietic lympho-myeloid progenitors to provide an unbiased assessment of transcriptional state and correlate this with function within LMPP, GMP and MLP. We sequenced 320 single cells that passed quality control from two different donors (157 and 163 from each donor; 91 LMPP, 110 MLP and 119 GMP). We performed clustering on cells from each donor and used the combined gene set variable in both donors to identify 3 clusters in both donors (Fig 6b and Supplementary Fig. 6d). The majority of the cells in cluster 1 were MLP and the majority of cluster 3 cells were GMP, while cluster 2 was comprised of LMPP and GMP cells (Fig. 6c and Supplementary Fig. 6e). Cluster 1 showed high expression of lymphoid-affiliated genes such as MME (CD10), JCHAIN (immunoglobulin J chain) and the transporter ABCA1. Conversely, cluster 3 showed increased expression of myeloid genes including CPA3 (Carboxypeptidase A3), MPO (Myeloperoxidase) and VIM (Vimentin) while cluster 2 showed a mixed transcriptional signature and increased expression of hematopoietic progenitor gene KIT. Principal component analysis on the single cell RNA sequencing data revealed a transcriptional
continuum of LMPP, MLP and GMP populations where PC1 captured the separation of cells into the three clusters (Fig 6di-dii). Identical analysis on the second donor revealed a similar structure to the data (Supplementary Fig. 6fi-fii).

**Further purification of LMPP and GMP**

Both the transcriptional and functional heterogeneity within LMPP, MLP and GMP indicate that current FACS based cell surface marker sorting does not purify different functional cells sufficiently. Therefore, we correlated surface marker expression and function within the LMPP and GMP populations, as these populations showed the greatest functional heterogeneity. Analysis of FACS indexing data showed that single LMPP cells with functional lymphoid output had significantly higher expression of CD10 and CD45RA compared to those with myeloid and lympho-myeloid output (Fig. 7a-b; CD10: Ly vs Ly-My p=0.0052, Ly vs My p=0.027; CD45RA: Ly vs Ly-My p=4.8x10^-6, Ly vs My p=0.0027, Wilcoxon rank sum test). This was further confirmed by the higher CD10 mRNA expression in single LMPPs in lymphoid-biased cluster 1, compared to myeloid-biased cluster 3 (Supplementary Fig. 7a).

Using this information, we developed new LMPP FACS-sorting strategies to purify two LMPP subpopulations (Supplementary Fig. 7b – methods for detailed explanation). In the first case, we aimed to purify a LMPP sub-population with higher CD10 and CD45RA expression to prospectively purify out the myeloid potential. We call this population LMPP^ly or LMPP lymphoid as the prediction was it would have mainly lymphoid potential. In the second case, we aimed to purify LMPP sub-population with lower CD45RA and CD10 expression to maximize retained myeloid potential. We termed this population LMPP^mix as it was predicted to have mixed lymphoid, myeloid and lympho-myeloid potential.

When we employed these sorting schemes, 26% of total LMPP were LMPP^ly and 27% were LMPP^mix (Fig. 7c). Functionally, LMPP^ly were intermediate between LMPP and MLP. LMPP^ly had significantly lower cloning efficiency compared to LMPP and LMPP^mix but significantly higher than MLP (Fig. 7d and Supplementary Table 9; Fisher’s exact test p<0.0001 for all comparisons). As expected LMPP^ly cells are lymphoid progenitors (Fig. 7eII). They have
virtually no myeloid potential (Fig. 7eii). There was significantly lower myeloid potential in LMPP\textsuperscript{ly} compared to LMPP and LMPP\textsuperscript{mix} (Fig. 7eii and Supplementary Table 9; Fisher’s exact test p=0.0496 and p=0.0280 respectively). There was a very small residual (1.6%) lymphoid-myeloid potential (Fig. 7i, purple hatched bars). In contrast, LMPP\textsuperscript{mix} cells retain virtually all of the myeloid potential (Fig. 7eii-h and Supplementary Table 9) and most of the lympho-myeloid potential (Fig. 7i, purple hatched bars). This was confirmed using a second in vitro culture system (SF7b) condition we used previously (Fig. 7j, Supplementary Fig. 7c-g and Supplementary Table 9).

We also devised new FACS purification strategy to remove lymphoid potential from the GMP. Significantly higher CD38 protein expression in GMPs with myeloid only output was observed compared to GMPs with lympho-myeloid or lymphoid output (p=1.57x10\textsuperscript{-11} and p=1.6x10\textsuperscript{-8} respectively, Wilcoxon rank sum test) (Fig. 7j-k). The maximum normalized CD38 level of GMPs with myeloid output was 2.80, whereas for lympho-myeloid output it was 2.43 (Fig. 7k). Concordantly, CD38 expression in GMPs measured by single cell RNA sequencing showed that cluster 3 (with the highest myeloid potential) had significantly higher CD38 expression (Supplementary Fig. 7h). Next, we showed a significant positive correlation between CD38 expression and myeloid (MPO) gene expression but a negative correlation between C38 and lymphoid gene (MME and SELL) expression in the single cell qRT-PCR data (p=2.2x10\textsuperscript{-16}, p=0.53 (MPO), p=7.1x10\textsuperscript{-5}, p=-0.22 (MME), p=1.3x10\textsuperscript{-5}, p=-0.24 (SELL), Spearman’s rank correlation coefficient, Supplementary Fig. 7i).

Based on this observation, we devised a novel sorting strategy to purify a GMP subpopulation to eliminate lymphoid potential (retaining myeloid potential) based on CD38 expression (Supplementary Fig. 7j – see methods for details). In this case, we studied the entire Lin\textsuperscript{-}CD34\textsuperscript{+} population and divided it into CD38\textsuperscript{hi} (44% of CD38\textsuperscript{+}), CD38\textsuperscript{lo} (15% of CD38\textsuperscript{-}) and CD38\textsuperscript{mid} cells (the area between the two new gates) that had not been previously functionally tested (Fig. 7i). CD38\textsuperscript{hi}, CD38\textsuperscript{mid} and CD38\textsuperscript{lo} cells were further purified to isolate GMP CD38\textsuperscript{hi}, CD38\textsuperscript{mid} (CD38\textsuperscript{mid}CD45RA\textsuperscript{-}CD10\textsuperscript{-}) and LMPP CD38\textsuperscript{lo}. Due to the low
abundance of LMPP CD38\textsuperscript{hi} cells (1 in 10\textsuperscript{8} MNCs; only 34 cells could be sorted from 4 CB donors) no conclusions could be reached about their functional potential and this population was excluded from further analysis. In vitro function of single GMP CD38\textsuperscript{hi} (279 cells) and CD38\textsuperscript{mid} (693 cells) was compared to conventionally defined LMPP (1136 cells) and GMP cells (1622). The cloning efficiencies of the 4 populations varied between 55\% (GMP CD38\textsuperscript{hi}) to 73\% (CD38\textsuperscript{mid}) (Fig. 7m). As expected all four populations mainly consisted of uni-lineage progenitors (63-72\%). Compared to conventionally purified GMP, GMP CD38\textsuperscript{hi} had no uni-lineage B cell potential and significantly less B cell potential in bi-lineage and multi-lineage outputs. GMP CD38\textsuperscript{hi} cells were also significantly depleted of NK and combined lympho-myeloid potential (Fig. 7n-r and Supplementary Table 9). Thus, GMP CD38\textsuperscript{hi} is principally a myeloid progenitor composed of heterogeneous uni-, bi- and very few multi-lineage progenitors. CD38\textsuperscript{mid} is a more lineage-balanced population with less lymphoid potential (B cell) and more GM potential than the LMPP, and more lymphoid potential and less myeloid potential (principally G) than the GMP (Supplementary Table 9). In summary, these two new revised FACS sorting schemes allow purification of more homogeneous populations within existing LMPP and GMP compartments.

Discussion

Though several human lympho-myeloid progenitors have been identified, five important issues have prevented a clear understanding of the relationship between their functional potential and transcriptional programs. First, lympho-myeloid progenitor populations, like other stem/progenitor populations, have been isolated using cell surface markers based on historical precedent rather than markers that purify to functional homogeneity. Second, prospectively isolated lympho-myeloid progenitor populations have never previously been systematically compared in the same experiments. Third, it is unclear if early progenitor populations downstream of HSC contain only unipotent cells\textsuperscript{16, 25} or also contain bi- and multipotential progenitors, more extensively studied in the mouse\textsuperscript{5, 14, 17, 18, 26, 39} than human\textsuperscript{21, 22, 23, 24}. Fourth, functional assays demonstrate potential rather than actual cell fate in vivo.
during steady state conditions. Finally, a failure to register a potential in a functional test may reflect the inadequacy of the assay rather than the potential, or indeed fate, of the cell in vivo. Thus, there is uncertainty about how distinct the differently identified progenitors are; and if distinct, what their relative functional potential and transcriptional programs are at a population and single cell level.

Here, we report on the prospective separation and direct comparison of three freshly isolated Lin−CD34+ lympho-myeloid progenitors: the LMPP, MLP and GMP. Within the Lin−CD34+ compartment we defined the LMPP as CD38−CD90loCD45RA22, unlike the CD62Lhi LMPP23, which is CD38+CD45RA+ and overlaps with the GMP. The GMP contains both CD62Lhi cells and cells with lower CD62L expression (data not shown). The MLP, on the other hand is also CD38′CD90′CD45RA+ but when first identified was reported as CD10−24 unlike the LMPP which is CD10−22. However, a recent report has suggested that there are CD10− MLP populations40, which may overlap with the LMPP. Taken together, what is consistent is that lympho-myeloid progenitors are in the Lin−CD34+CD45RA−CD90− compartment and can be either CD38+ or CD38− and CD10− or CD10+. This led us to directly compare the LMPP, MLP and GMP as defined in Fig. 1, as it would capture all these populations.

Our data show that though all three populations contain both lymphoid and myeloid progenitors, they are functionally different. We exhaustively tested function of 4598 single LMPPs, MLPs and GMPs as well as populations of these progenitors in three different and efficient, new and published in vitro culture conditions. In addition, we also employed a new, more efficient in vivo assay to corroborate in vitro findings. The sum of these functional observations demonstrate that the GMP is primarily a myeloid progenitor in vitro and in vivo with residual B-cell and NK cell progenitor potential. The residual lymphoid potential can be virtually eliminated by purifying GMP within a gate containing the highest 44% of CD38-expressing cells. The MLP, conversely, is primarily a B-cell, NK cell and T-cell progenitor with residual monocyte progenitor activity in vitro. This is mirrored in vivo where the MLP produces lymphoid cells; the minor CD33+ cell output does not express mature myeloid cell markers.
The LMPP has progenitors with both lymphoid and myeloid potential. In vitro it is predominantly a lymphoid progenitor though 12-27% of progenitors are myeloid or lympho-myeloid in the different culture conditions. We propose a new FACS purification scheme that divides the LMPP into two populations based on CD10 and CD45RA; one which is almost entirely lymphoid and the other that captures most of the myeloid potential. Interestingly, however, the LMPP produces mainly myeloid cells in vivo, suggesting that the humanized ossicles may be particularly efficient at promoting human myelopoiesis, unlike naive NSG mice, which better support human lymphopoiesis.

RNA-seq analysis at a population and single cell level confirms the functional observations with primarily myeloid lineage priming in the GMP, lymphoid lineage priming in the MLP and a more mixed lympho-myeloid signature in the LMPP. Expression of lineage-affiliated transcriptional regulators and cytokine receptors is in accord with this. Single cell transcriptional analysis is revealing, showing the three populations form a transcriptional continuum with MLP cells arcing from a lymphoid pole and GMP from myeloid pole to intersect with the LMPP cells positioned at the arc apex.

Four final key observations are worth noting. First, 24 different lineage-affiliated potentials were detected in lympho-myeloid progenitors (Fig. 3o). This is likely to be an underestimate, as we did not test for eosinophil, mast cell, basophil and dendritic cell function. Second, though the majority of progenitors were uni-lineage, bi-potential and multi-lineage output was seen (up to 39% and 13%, respectively, of cells that expand in vitro). This suggests that the lympho-myeloid lineage decisions may occur at multiple levels in the hemopoietic hierarchy namely, HSC1, 2, 3, MPP5, 14 and presumably more mature LMPP17, 18, 26, MLP and GMP39 populations. Furthermore, within the LMPP and GMP true lympho-myeloid progenitors are likely to be rare (up to 10-14% of cells that expand in vitro) and concentrated in the LMPP. These observations suggest that an alternative hierarchy of hemopoiesis may exist (Fig. 8), different to that previously been proposed (Fig. 1a). In this model the hierarchical relationships between lineage-biased HSC, MPP and the progenitors we have studied still
remain to be established. One key question is whether the diverse lineage-affiliated progenitors have stably different functions or whether there is plasticity such that their functional output may be stochastically determined or variably instructed. Further single cell functional analysis on potentially more functionally pure populations will be required and more detailed fate mapping in mice will be informative. Third, the rarity of multi-potential and bi-potential progenitors is critical to bear in mind. In this regard, it is first important to appreciate that the LMPP and MLP themselves are rare (~2/10^5 BM MNCs and ~1/10^4 CB MNCs). Next within the LMPP, MLP and GMP bi- and multi-potential progenitors are only a fraction of all the cells. This is relevant as single cell RNA-Seq programs will have to sequence very large numbers of MNCs to provide adequate representation of rare progenitors especially those subsets with bi- or multi-lineage potential. It is for this reason that we studied extremely large numbers of single cells to get robust information on rare bi- and multi-lineage potential in progenitors. Fourth, all our functional and transcriptional analyses were performed on CB. Though the LMPP, MLP and GMP have not been studied in detail through development, quantitative differences in multi- versus uni-lineage output have been documented in in vitro assays between fetal liver and bone marrow in the broad CD34^+CD38^- populations. Finally, the rarity of the LMPP is also remarkable given that acute myeloid leukemia (AML) stem cells are often arrested at an LMPP-like stage where they can comprise up to 80% of MNCs. Given this, understanding how normal LMPPs differentiate may provide insight into novel differentiating therapies for AML.

**Methods**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

**Acknowledgements**

We thank Juan Carlos Zuniga-Pflucker for OP-9 DL4 cells used for initial experiments. We would like to acknowledge funding from the MRC (MHU Award G1000729, MRC Disease Team Award 4050189188), CRUK (Program Grant to PV C7893/A12796, CRUK program grant to BG C1163/A21762), Bloodwise (Specialist Program 13001 and Project grant to
12019), an MRC PhD studentship (F.H. & Z.A.), The MRC Single Cell Award (MR/M00919X/1) to the WIMM and the Oxford Partnership Comprehensive Biomedical Research Centre (NIHR BRC Funding scheme). We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z) for generation of sequencing data. R.M. was supported by National Institutes of Health grants R01CA188055 and U01HL099999, New York Stem Cell Foundation Robertson Investigator and Leukemia and Lymphoma Society Scholar Award. A.R. was supported by an Erwin-Schroedinger Research fellowship from the Austrian Science Fund (FWF).

Author Contributions
D.K., B.S., Z.A., and P.V. designed the experiments; D.K., B.S., Z.A., A.R., M.S., L.Q., and N.G. performed experiments and analyzed data; F.H., G.O., Z.A., E.R. and S.T. performed bioinformatics and statistical analysis; J.D. and B.U. prepared samples; J.C., E.S., F.P., R.M., C.P. and B.G., provided reagents and materials; D.K., B.S. and P.V wrote the paper; All authors edited the manuscript.

Competing Financial Interests
Nil

References:


**Figure legends**

**Figure 1. Immunophenotypic separation of eight distinct human hematopoietic stem/progenitor cell (HSPC) populations and their frequencies in CB and bone marrow (BM).**

a) Lineage output of lympho-myeloid progenitors (a1) and models of human hemopoietic differentiation and lineage diversification (a1i-iii). Dashed arrows - hierarchical relationships not experimentally validated.

b) Immunophenotypic characterization of HSPC populations studied.

c) Representative plot of FACS-sorting of human CB HSPCs. Population frequencies shown within plots are the mean from 44 biological replicates calculated as a percentage of Lin−CD34+ compartment.

d) Mean frequency (± SD) of immunophenotypically defined HSPC populations as a percentage of Lin−CD34+ fraction in human CB (N=44) and BM (N=7) using analysis gates.

e) Mean frequency (± SD) of immunophenotypically defined HSPC compartments as a percentage of Lin−CD34+ fraction of human CB (N=44) using FACS sorting gates.

HSC, hemopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; MLP, multi-lymphoid progenitor; CMP, common myeloid progenitor;
Figure 2. Human CB lympho-myeloid populations have distinct functional potential in vitro.

a) Cloning efficiency and lineage affiliation of myeloid and erythroid colonies in a CFU assay (150 CB HSPCs plated). Error bars are ±1SD. N=5. Abbreviations of HSPC populations as in Figure 1. CFU-mix, mixed erythro-myeloid colony; CFU-M, monocyte/macrophage colony; CFU-G, granulocyte colony; CFU-GM, granulocyte and monocyte/macrophage colony; E, erythroid colony (both BFU-E and CFU-E).

b) Upper panel: Morphology of May-Giemsa Grunwald stained cells from the CFU assay. Bar size, 10 mm. Lower panel: FACS plots of cells harvested from indicated colony types. Antibodies used are indicated.

c) Lineage output after culturing 150 LMPP, MLP and GMP cells for 1, 2 or 3 weeks on MS-5 stroma with SCF, G-CSF, FLT3L, IL15, IL2 and DuP-697. Data represents mean from 3 biological replicates ± SD. FACS plots for two week cultures are shown in Supplementary Figure 2a.

d) Gene expression analysis of FACS-purified output cells from (c).

e) T-cell output after culturing LMPP, MLP and GMP cells in bulk for 5 or 7 weeks on OP9-hDL1 stroma with SCF, FLT3L and IL7. Data represents mean from 5 biological replicates ± SD, calculated as a percentage of human CD45+ cells. DN, CD7⁺CD1a⁺ CD4⁻CD8⁻; DP (double positive), CD7⁺CD1a⁺CD4⁺CD8⁺; ISP (immature single positive), CD7⁺CD1a⁺CD4⁺CD8⁻; CD7⁺CD1a⁺CD4⁻CD8⁺. FACS plots for five week cultures are shown in Supplementary Figure 2b.

f) Gene expression analysis of cDNA from FACS-purified output cells from the T-cell assay and control mature non T-cells. E, erythroid cells; G, granulocytes; M monocytes. E, G and M cells were obtained from sorting cells from E, G and M colonies.
Figure 3. CB LMPP and GMP are lympho-myeloid progenitors, while MLP is mainly a lymphoid progenitor in liquid culture quantitative in vitro assays.

a) Experimental strategy for sorting, culture and analysis for quantitative in vitro assays in SGF15/2 condition. Similar strategy was used for other culture conditions.

b) Frequencies of lineage outputs from LDA, shown as “1 in X cells can give rise to”. ND – not detected.

c) Cloning efficiency of the single cell LMPP, MLP and GMP in SGF15/2 condition. Number of cells producing output over total number of cells cultured is shown. Significance was defined using Fisher’s exact test.

d) Cloning efficiency of lymphoid lineages (di) and myeloid lineages (dii) of single cell LMPP, MLP and GMP in SGF15/2 condition. Bars indicate total cloning efficiency; filled portion indicates the proportion of lymphoid potential (lymphoid plus mixed clones) or myeloid potential (myeloid plus mixed clones). Mean ± SD is shown. Significance was defined using two-tailed students unpaired t-test.

e) Single-, f) bi- and g) multi-lineage outputs from single cells, presented as a percentage of the wells producing output in SGF15/2 condition.

h) Lymphoid (Ly), myeloid (1 My) and lympho-myeloid (Ly-My) outputs presented as a percentage of all plated single LMPP, MLP and GMP cells in SGF15/2 condition.

i) Cloning efficiency of the single cell LMPP, MLP and GMP in SF7b/Dox condition. Significance was defined using Fisher’s exact test.

j) Cloning efficiency of lymphoid lineages (ji) and myeloid lineages (jii) of single LMPP, MLP and GMP cultured in SF7b/Dox condition. Bars indicate total cloning efficiency; filled portion indicates the proportion of lymphoid potential (lymphoid plus mixed clones) or myeloid potential (myeloid plus mixed clones). Mean ± SD is shown. Significance was defined using two-tailed students unpaired t-test.

k) Single-, l) bi- and m) multi-lineage outputs from single cells, presented as a percentage of the positive plated wells in SF7b/Dox condition.
n) Lymphoid (Ly), myeloid (My) and lympho-myeloid (Ly-My) outputs presented as a percentage of all plated single cell LMPP, MLP and GMP wells in SF7b/Dox condition.

o) Summary of lineage outputs from single LMPP, MLP and GMP. For the LDA: data from 4 biological replicates. For the single cell assay in SGF15/2 condition: data from 22 biological replicates. For the single cell assay in SF7b/Dox condition: data from 3 biological replicates.

Figure 4. Human CB LMPP, MLP and GMP progenitors have distinct differentiation potential in vivo.

a) Experimental strategy for generation of human ossicles in NSG mice and human progenitor transplantation.

b) Percentage human engraftment 2 weeks after progenitor transplantation normalized to 1000 transplanted human progenitor cells.

c) Percentage B cells and myeloid cells within human CD45+/HLA-ABC+ cells.

d) Representative FACS plots of percentage human engraftment (CD45+HLA-ABC+), % B (CD19+) and myeloid cells (CD33+) and % CD14+ and CD15+ myeloid cells 2 weeks after transplantation. The frequencies shown are average from 11 biological replicates for LMPP, 3 from MLP, 6 for GMP.

e) Representative images of May- Grunwald-Giemsa stained CD15+, CD15+/CD14+ and CD14+ myeloid cells generated by LMPP 2 weeks after transplantation, N=2.

Figure 5. Distinct transcriptional patterns of human CB HSPC populations.

a) Hierarchical clustering of HSPC populations using all genes and 1000 bootstrap (bs) permutation analyses; “au” = approximate unbiased p-values indicate the probability that the cluster is supported by data. 100 = 100% probability that the cluster is correct. Height values measure dissimilarity, expressed as (1- [correlation co-efficient]). Correlation co-efficient of 1.0 = complete correlation.
b) PCA plots showing position of normal CB HSPC when using varying number of genes, using ANOVA (ranked by ANOVA p-value) and 300 most variant genes (bottom right). Percentage variance represented by each Principal Component (PC) is shown.

c) Loadings plot, showing the genes with the most extreme loadings scores for the PCA run with top 300 ANOVA genes (top) and 300 most variant genes (bottom).

d) Heatmap showing the expression of the top 300 ANOVA genes by the HSPC populations with hierarchical clustering on genes and populations. Clusters highlighted in yellow show distinct expression patterns across HSPC populations. Expression values are normalized per gene (by row) and shown as a minimum-maximum scale.

e) Summary of the number of all differentially expressed genes between HSPC populations.

f-g) Heatmaps showing the expression of (fi-iii) top 50 most differentially expressed genes (ranked by p value, adjusted for multiple testing) between LMPP, MLP and GMP and the rest of HSPC populations and (g) transcription factors differentially expressed across HSPC populations. Genes affiliated with the lymphoid or myeloid lineages are color-coded (lymphoid: orange, myeloid: green) and genes associated with immune function are labeled in black. Expression values are normalized per gene (by row) and shown as a minimum-maximum scale.

RNA seq data come from 3-4 independent replicates.

**Figure 6. Transcriptional heterogeneity of CB lympho-myeloid progenitor cells from single cell RNA-sequencing.**

a) Experimental scheme used to combine single cell functional analysis, single cell RNA sequencing and single cell qRT-PCR using FACS index data.

b) Heatmap showing clustering of single LMPP, GMP and MLP using the 55 most highly and variably expressed genes between clusters. The heat map shows clustering from one of two donors used. Data from the other donor is in **Supplementary Fig. 6d**. Log-normalised gene expression (rows) for each single cell (columns) is shown. 3 clusters are detected.
c) The contribution of LMPP, GMP and MLP to each of the 3 clusters identified in (b) is shown.

d) PCA plot colored in cell type (di) or by cluster membership (dii).

**Figure 7. New FACS sorting strategies to purify functional potential within CB LMPP and GMP compartments.**

a) Logicle transformed CD10 (ai) and CD45RA (aii) surface marker levels in LMPPs, grouped by functional output. Ly - uni-lymphoid (B or NK) output (B or NK) or bi-lymphoid output (B+NK), My - uni-myeloid output (M or G) or bi-myeloid output (M+G), Ly-My - lympho-myeloid output. N=2 biological replicates.

b) CD10 and CD45RA expression levels, measured by FACS, in LMPPs colored by output from functional assays. Logicle-transformed data are from 2 biological replicates.

c) Revised sorting strategy based on CD10 and CD45RA expression levels defined by bioinformatic analyses. LMPP\(^{ly}\) and LMPP\(^{mix}\) single cells were index sorted for in vitro functional assay. Representative plots from 6 biological replicates.

d) Cloning efficiency of single MLP, LMPP, LMPP\(^{ly}\), LMPP\(^{mix}\) and GMP in SGF15/2 condition. Significance was defined using Fisher’s exact test.

e) Cloning efficiency of lymphoid lineages (ei) and myeloid lineages (eii) of single cell MLP, LMPP, LMPP\(^{ly}\), LMPP\(^{mix}\) and GMP in SGF15/2 condition. Bars indicate total cloning efficiency; filled portion indicates the proportion of lymphoid potential (lymphoid plus mixed clones) or myeloid potential (myeloid plus mixed clones). Mean ± SD is shown. Significance was defined using two-tailed students unpaired t-test.

f) Single- g) bi- and h) multi-lineage outputs from single cells in SGF15/2 condition, presented as % of the plated wells with output.

i) Lymphoid (Ly), myeloid (My) and lympho-myeloid (Ly-My) outputs presented as a percentage of all plated single MLP, LMPP, LMPP\(^{ly}\), LMPP\(^{mix}\) and GMP cells in SGF15/2 condition.
j) Lymphoid (Ly), myeloid (My) and lympho-myeloid (Ly-My) outputs presented as a percentage of all plated single MLP, LMPP, LMPP\textsuperscript{ly}, LMPP\textsuperscript{mix} and GMP cells in SF7b condition.

k) Logicle transformed CD38 surface marker levels in GMPs, grouped by functional output. N=5 biological replicates.

l) CD38 and CD34 levels in GMPs colored by output from functional assays. Data are from 5 biological replicates.

m) Revised sorting strategy, based on CD38 expression levels defined by bioinformatic analysis. GMP CD38\textsuperscript{hi} (44\% of CD38\textsuperscript{+}), CD38\textsuperscript{mid} (CD45RA\textsuperscript{-}CD10\textsuperscript{-}) and LMPP CD38\textsuperscript{lo} (15\% of CD38\textsuperscript{-}) single cells were index sorted for in vitro functional assay. Representative plots from 4 biological replicates.

n) Cloning efficiency of the single GMP CD38\textsuperscript{hi}, GMP, CD38\textsuperscript{mid} and LMPP. Significance was defined using Fisher’s exact test.

o) Cloning efficiency of lymphoid lineages (ni) and myeloid lineages (nii) of single cell GMP CD38\textsuperscript{hi}, GMP, CD38\textsuperscript{mid} and LMPP. Bars indicate total cloning efficiency; filled portion indicates the proportion of lymphoid potential (lymphoid plus mixed clones) or myeloid potential (myeloid plus mixed clones). Mean ± SD is shown. Significance was defined using two-tailed students unpaired t-test.

p) Single-, q) bi- and r) multi-lineage outputs from single cells, presented as \% of the plated wells with output.

s) Lymphoid (Ly), myeloid (My) and lympho-myeloid (Ly-My) outputs presented as a percentage of all plated single GMP CD38\textsuperscript{hi}, GMP, CD38\textsuperscript{mid} and LMPP cells.

For functional assay based on the revised LMPP sorting strategy (d-j) data are from 6 biological replicates (SGF15/2 condition) and 6 biological replicates (SF7b condition). For functional assays based on revised GMP strategy (n-s), data from 4 biological replicates. For single cell functional assay (SGF15/2 condition) for LMPP, MLP and GMP controls (d-i and n-s) data from 22 biological replicates (the same shown in Fig. 3c-h). For single cell functional
assay (SF7b condition) for LMPP, MLP and GMP controls data from 9 biological replicates (including 3 biological replicates used for Fig. 3i-n).

**Figure 8. Model of human lympho-myeloid differentiation**

Multiple, rare, functionally distinct bi- and multi-lineage lympho-myeloid progenitors (LMP) could differentiate either directly from hematopoietic stem cells (HSC) or multi-potent progenitors (MPP). We have used the term LMP, rather than LMPP (lymphoid primed multi-potential progenitor) to describe these lympho-myeloid progenitors, as not all LMP will be lymphoid biased. Multi-lineage LMPs are rare. Lymphoid only and myeloid only progenitors are shown below. Bi-lineage progenitors are more frequent and uni-potent progenitors are most common. The hierarchical relationships between LMP and lymphoid and myeloid only bi- and uni-lineage progenitors remain to be determined. This model also leaves open the question of when commitment to either lymphoid or myeloid fate occurs. Erythro-megakaryocytic differentiation leading to a MEP (megakaryocyte-erythroid progenitor), erythroid (BFU-E) and megakaryocytic progenitor (MkP) could occur either directly from HSC or MPP.

**Online Methods**

**Normal and patient samples collection**

BM or CB samples from normal donors were obtained with informed consent (UK protocol MREC 06/Q1606/ or Administrative Panel on Human Subjects Research Institutional Review Board-approved protocols Stanford IRB no. 18329, no. 6453, and no. 5637). Fresh CB samples were purchased from NHS Cord Blood Bank, UK or from New York Blood Center. They were processed within 16-34h after collection. Mononuclear cells were isolated and CD34⁺ fraction was separated as described¹. Fresh or frozen BM MNCs or CD34⁺ fractions were used. Human BM stromal cell were obtained from samples according to Medical University of Graz Ethikkommission (Institutional Review Board-approved protocol, MUG Graz IRB no. 19-252). BM mesenchymal stromal cells (MSCs) were isolated and expanded as described².
FACS sorting of HSPC populations

All antibodies used for FACS sorting and immunophenotyping are listed in Supplementary Table 1. CB or BM CD34⁺ enriched fraction was lineage depleted by staining with purified anti-human CD2, CD3, CD4, CD7, CD8a, CD11b, CD14, CD19, CD20, CD56, CD235a followed by Qdot 605 conjugated goat F(ab′)2 anti-mouse IgG (H+L). Cells were also stained with anti-human CD38-FITC, CD45RA-PE or -BV650, CD123-PE Cy7, CD90-biotin, CD34-PerCP and CD10-APC. Finally, cells were incubated with streptavidin-conjugated APC-eF780 and Hoechst 33258 (Invitrogen, Loughborough UK; final concentration: 1 µg/ml). For humanized ossicle xenotransplantation assay CD34⁺ CB was stained with the same panel of anti-human lineage antibodies and anti-CD16. All lineage-antibodies were PE Cy5-conjugated. Cells were then stained with CD38-PE Cy7, CD90-FITC, CD123-PE, CD34-APC, CD10-APC Cy7, CD45RA-BV650 and propidium-iodide (Thermo Fisher, Waltham MA; final concentration: 1 µg/ml). Unstained, single stained and Fluorescence Minus One (FMO) controls were used to determine background staining and compensation in each channel (Supplementary Fig. 1a). Single stained controls used anti-mouse compensation particle set (BD, Oxford UK). CB cells were sorted with average purity 99% for in vitro and RNA assays (Supplementary Fig. 1b) and 96% for humanized ossicle xenotransplantation (Supplementary Fig. 3b). Prior to single-cell sorts, single fluorescent beads were deposited directly to a 96-well plate to establish accuracy of single cell deposition (>99%). Sorting was performed on BD Aria III or BD Fusion and FACS analysis was done on LSR Fortessa X20. Data analysis was performed using Diva v8.1 or FlowJo v10.0.06 and v10.0.07r2.

Index sorting for functional and transcriptional analyses

For index sorting we saved information on the following parameters: FSC, SSC, Hoechst and expression of Lineage markers, CD34, CD38, CD45RA, CD10, CD90 and CD123 for each single cell. For 919 index sorted single cells we tested expression of 96 genes qRT-PCR (Supplementary Fig. 6); 74 passed QC. Separately, we performed single cell index sorting and single cell in vitro functional assays on 3458 single cells (From Fig. 3, Supplementary Fig. 3, Figure 7, Supplementary Fig. 7). In separate experiments we index sorted 320
single cells for single cell RNA seq (Fig. 6). Using common “position of the cells” in FACS plots we can then map functional potential (i.e. lymphoid, myeloid or lympho-myeloid) to gene expression and cell surface marker expression and forward/side scatter. To purify LMPP^b and LMPP^mix the thresholds were defined based on maximum CD10 and CD45RA expression of LMPPs with myeloid output. To purify GMP CD38^hi thresholds were set using the maximum normalized CD38 level of GMPs with myeloid output and for lympho-myeloid output.

**In vitro lympho-myeloid differentiation assays (bulk, single cell, LDA)**

For population analysis, MS-5 cells were seeded on a 24-well plate coated with 0.1% gelatin at a density of 2x10^4 cells per well in α-MEM medium (Gibco/Thermo Fisher Scientific Loughborough UK) supplemented with 10% FBS (Hyclone, GE Healthcare, SH30070.03 Amersham Hatfield, UK), 1% Penicillin-Streptomycin, 1% L-Glutamine, 10^-7M DuP-697 (Cayman Chemical, Ann Arbor, USA), 20 ng/ml SCF, 10 ng/ml G-CSF, 10 ng/ml FLT3L, 10 ng/ml IL15 and 10 ng/ml IL2 (Peprotech London UK, **SGF15/2 condition**). 24h after plating of MS-5 cells, 150 highly purified LMPP, MLP or GMP were deposited in each well. Medium was half-changed every week. Harvested cells were FACS analyzed at week 1, 2 and 3.

LDA was performed by sorting LMPP, MLP or GMP cells at different cell doses (1, 2, 5, 10 and 20 cells) from 4 different CB samples into 96-well plates pre-plated with 2500 MS-5 cells per well with 100 µl of medium without cytokines. Immediately after sorting 100 µl of 2x SGF15/2 medium was added to each well. Medium was half-changed every week. A total of 833 LMPP, 789 MLP and 1252 GMP cells from 4 different CB samples were analyzed for the LDA at week 2 – 2.5 (**Supplementary Table 2a**). Frequency calculations were performed using L-Calc software (Stem Cell Technologies) and independently verified by ELDA software ([http://bioinf.wehi.edu.au/software/elda/](http://bioinf.wehi.edu.au/software/elda/)). The LDA plots were generated using R with lines representing the estimates calculated by ELDA software.

For single cell analysis single LMPP, MLP and GMP cells were deposited into 96-well plates pre-plated with 2500 MS-5 cells per well with 100 µl of medium without cytokines. Medium with 2x cytokines was added to each well after sorting. Medium was half-changed every
week. After culture for 2-2.5 weeks FACS analysis was performed and wells with more than 15 human CD15+, CD14+, CD56+ or CD19+ cells were scored positive (details in Supplementary Table 2b). To compare with previous published conditions, single cell LMPP, MLP and GMPs were cultured for 4 weeks on MS-5 stroma in H5100 medium (StemCell Technologies Cambridge UK) supplemented with 100 ng/ml SCF, 20 ng/ml IL-7, 50 ng/ml TPO, 10 ng/ml IL-2, 20 ng/ml GM-CSF, 20 ng/ml G-CSF and 10 ng/ml M-CSF (all from Peprotech, S7T2GM/G/M condition) and analyzed by FACS.

To read lineage readouts for all in vitro lympho-myeloid differentiation assays, harvested cells were stained with anti-human CD15-FITC, CD14-PE, CD19-PE Cy7, CD56-APC or -PE Cy5, CD45-APC Cy7 and in some cases with CD34-BV605.

**In vitro T-cell differentiation assay**

OP9-hDL1 cells were seeded on a 24-well plate coated with 0.1% gelatin at a density of 2x10^4 cells per well in freshly prepared α-MEM medium (Gibco/Thermo Fisher Scientific, 12000-063) with 20% heat-inactivated FBS (Hyclone, GE Healthcare, SH30070.03 Amersham Hatfield, UK), 1% Penicillin-Streptomycin, 1% L-Glutamine, 10 ng/ml SCF, 5 ng/ml FLT3L and 5 ng/ml IL7 (Peprotech, London, UK, SF7a condition). 24h after OP9-hDL1 cell plating, 150 highly purified LMPP, MLP or GMP cells were deposited in each well. Cells were dissociated from wells and transferred to new plates with fresh OP9-hDL1 cells weekly. Harvested cells were FACS analyzed at week 4, 5 and 7. Cells were stained with anti-human CD7-FITC, CD1a-PE, CD8-PE Cy7, CD4-APC and CD45-APC Cy7.

**In vitro combined T-lympho-myeloid differentiation assay**

MS5-hDL1^IND^100 cells were seeded on 96-well plates coated with 0.1% gelatin at a density of 2500 cells per well in 100 µl freshly prepared α-MEM medium (Gibco/Thermo Fisher Scientific, Loughborough UK) supplemented with 20% FBS (Hyclone, GE Healthcare, SH30070.03HI, Amersham Hatfield, UK), 1% Penicillin-Streptomycin, 1% L-Glutamine. 24h after plating of MS5-hDL1^IND^ cells, single cell LMPP, MLP or GMP cells were deposited into each well and cultured in the presence of 20nM Insulin (Sigma-Aldrich, St Louis, MO), 50
ng/ml SCF, 20 ng/ml FLT3L and 10ng/ml IL7 (Peprotech London UK, **SF7b condition**). Fresh medium was added every week.

Cells were harvested at 21 days and split into two, half of them were used for FACS analysis and the remaining half were re-seeded on MS5-hDL1\textsuperscript{IND}100 cells and cultured in SF7b/Dox condition with doxycycline (1 µg/ml). Medium was half-changed twice every week. Fresh doxycycline was added to the cultures 3 times a week. At 42 days cells were harvested and FACS analysis was performed. At 21 days wells with more than 8 human CD15\textsuperscript{+}, CD14\textsuperscript{+}, CD56\textsuperscript{+} or CD19\textsuperscript{+} cells were scored positive. At 42 days FACS analysis using CD1a, CD7, CD4 and CD8 antibodies was performed and wells with more than 8 CD7\textsuperscript{+} cells were scored positive for T cells.

**Colony Forming Unit assays**

Colony formation was tested as before\textsuperscript{5}. Colony identity was confirmed morphologically after cytospin (medium acceleration, 800 rpm 5 min May-Grunwald Giemsa stain (Sigma, Poole UK) and by FACS with anti-human CD15-FITC, CD14-PE, CD235a-PE Cy5.

**Humanized ossicle xenotransplantation assay**

Protocol was performed as previously described\textsuperscript{2}. In brief, in vitro expanded human BM-MSCs were harvested, resuspended in 60 µl of pooled human platelet lysate (pHPL) and admixed with 240 µl of matrigel-equivalent matrix. The whole matrix-cell mixtures were injected subcutaneously to generate humanized ossicle niches. 8-10 weeks post BM-MSC application transplants were evaluated for bone and marrow formation. Mice with established humanized ossicle niches were conditioned with 200 rad of irradiation 12-24 hours prior to transplantation. Different numbers of LMPP, MLP and GMP cells from at least 3 different CB donors (Supplementary Fig. 4c) were transplanted in total volume of 20 µl by direct intraossicle injection. Experiments were performed in accordance with a protocol approved by Stanford’s Administrative Panel on Laboratory Animal Care (no. 22264) and in adherence to the US National Institutes of Health’s Guide for the Care and Use of Laboratory Animals. Normal multi-lineage engraftment was assessed 1-2 weeks after transplantation and defined

33
by the presence of myeloid cells (CD33⁺) and B-cells (CD19⁺) among engrafted human CD45⁺HLA-ABC⁺ cells. Engrafted mice were antibody stained with CD14-PE or -APC Cy7, CD15-FITC, HLA-ABC-FITC or -PB, CD19-APC, CD33-PE, CD45-V450.

**RNA sequencing of bulk HSPC populations**

100 highly purified HSPCs from normal CB samples were sorted directly into lysis buffer in RNAse inhibitor (Clontech St Germain-en-Laye France) and stored at -80°C before further processing. cDNA synthesis was done with Smarter Ultra low input RNA kit v1 (Clontech) as previously described⁷. Illumina libraries were generated using Nextera XT DNA sample preparation kit and Index Kit (Illumina Chesterford UK). Library size and quality were checked using Agilent High-Sensitivity DNA chip with Agilent Bioanalyser (Agilent Technologies Stockport UK). Concentration of indexed libraries was determined using Qubit High-Sensitivity DNA kit (Invitrogen Loughborough, UK). Libraries were pooled to a final concentration of 5-14 nM and were sequenced on an Illumina HiSeq 2000 single-end 50bp reads.

**Bulk and single cell gene expression analysis by Dynamic Arrays**

Gene expression analysis was performed as described¹. TaqMan assays (Applied Biosystems) are listed in Supplementary Table 2 and 8.

**Single cell RNA sequencing**

Single cell libraries for RNA sequencing were prepared using the Smart-seq2 protocol⁸, where 23 cycles were used for the cDNA library preamplification. Illumina Nextera XT DNA sample preparation kit and Index Kit (Illumina Chesterford UK) was used for cDNA tagmentation and indexing. ERCC RNA Spike-In Mix (Ambion) was added to the lysis mix at a final dilution of 1:80,000,000. Library size, quality and concentration were checked as done for the bulk RNA sequencing. Libraries were pooled to a final concentration of 7-28 nM and 78 to 95 single cell libraries were combined per pool. Sequencing was done on HiSeq4000 using 75bp paired-end reads. Each pool contained a library generated from an empty well.

**Bioinformatic analysis (bulk RNA seq, single cell Biomark and single cell RNA seq)**
For 50 bp single end bulk RNA sequencing, alignment to the hg38 reference genome was carried out using TopHat v2.0.10. Alignments were processed using Picard tools (http://picard.sourceforge.net/). We used R version 3.1.1 http://www.R-project.org. Sequencing reads were filtered for mapq 4 i.e. uniquely mapping reads. This gave a range of $15.1 \times 10^6$ to $56.2 \times 10^6$ aligned reads. The total number of genes expressed per sample was calculated as an rpkm>1. The number of expressed genes ranged from 7,707 to 11,350, with an average of 9,800. The count matrix was transformed to log2(cpm) scale and Principal Component Analysis was carried out. An ANOVA-like test was performed, using edgeR package for R, to identify differentially expressed genes between the populations. One biological replicate MPP population was excluded because when compared to the 3 remaining MPP biological replicates its global gene expression showed higher number of uniquely expressed genes and low correlation to the other three replicates. The genes were ranked by their significance (p-value adjusted for multiple testing) and different numbers of genes were used for PCA and hierarchical clustering of samples. Eigenvalues from PCA were calculated by using the square of the standard deviation of the principle components. Differential gene expression for one versus one and one versus all comparisons were calculated using edgeR. For gene signature generation a cut-off of logFC>1 was used and genes ranked based on p-value. Heatmaps and associated hierarchical clustering were generated using GENE-E software (Broad Institute) or using the R packages pvclust and heatmap.2 (gplots). MetaCore Pathway Map (Thomson Reuters, London UK) enrichment analysis was carried out on genes differentially expressed by each lympho-myeloid population versus all other populations (one versus all). A p-value cut-off of 0.05 was used to identify positively enriched pathway maps.

Analysis of single cell Biomark data was performed in R version 3.3.1 using data exported from the Fluidigm Data Collection software. For quality control, amplification curves with a Quality Score of <0.65 and any Ct values >27 were treated as undetected expression. Any cells where expression of both B2M and GAPDH housekeeping genes was not detected were removed from further analysis (N=7). An additional cell was removed as it had a high outlying
number of genes detected. Housekeeping gene ACTB was also measured in the assay, but unlike B2M and GAPDH did not show robust expression across the majority of cells and therefore was not used in further analysis. Normalized ΔCt values were calculated by subtracting the mean of Ct values for B2M and GAPDH in each cell, as previously described\textsuperscript{10}. Housekeeping genes were excluded from further analysis. Genes detected in <20 cells, with variance <1 across all cells or expressed in none of the MLP, GMP or LMPP 10 cell control samples assayed by qRT-PCR alongside single-cell samples were removed from downstream analysis. Post quality control data measured 74 genes in 919 single cells. Hierarchical clustering was performed on genes and cells by using the hclust function (stats package) with distance measure 1 – Spearman’s correlation and agglomeration method Ward.D2. The heatmap visualizing the clustering was plotted using the heatmap.2 function (gplots package). Cells were divided into three clusters using the cutree function (stats package) on the hierarchical clustering. A gene was classed as differentially expressed between two clusters if it satisfied two criteria: 1) the magnitude of the log2 fold change of mean ΔCt in each cluster was >1 and 2) the adjusted p-value (Benjamini & Hochberg correction for multiple testing) of 2-sided Wilcox test of ΔCt expression values between the two clusters was < 0.01. Diffusion maps\textsuperscript{11} were used for dimensionality reduction of the single cell gene expression data. This method was implemented using the DiffusionMap function from the destiny R package with Euclidean distance\textsuperscript{12,13}. Single cell RNA sequencing reads were aligned using G-SNAP\textsuperscript{14} and mapped reads were assigned to Ensembl genes (release 81\textsuperscript{15}) by using HTSeq\textsuperscript{16}. Cells with fewer than 500,000 reads mapping to nuclear genes, greater than 20% of mapped reads mapping to mitochondrial genes, greater than 20% of mapped reads mapping to External RNA Controls Consortium (ERCC) spike-ins or with expression of fewer than 750 different genes with at least 10 counts were removed from further analysis. ERCC spike-in controls identified genes exceeding technical variance\textsuperscript{17}. From donors 1 and 2, 163/166 and 157/249 cells passed quality control, respectively. Single cell profiles were normalized using the scran R package\textsuperscript{18} and variable genes were identified as having variation exceeding technical levels\textsuperscript{17}. Data
showed batch effects between different donors. The Seurat R package (https://github.com/satijalab/seurat) was then used to regress out plate effects from the sequencing data, and set more stringent thresholds for variable genes, leading to 1,605 variable genes in donor 1 and 1,273 variable genes in donor 2. Principal component analysis was performed using Seurat, and clusters found using the Seurat::FindClusters function on the first 10 principal components. Heatmaps display the top genes marking these clusters as identified by the Seurat::FindAllMarkers function and were visualized using the gplots::heatmap.2 function.

**Statistical analysis**

Frequency of populations in FACS plots gates is the mean of the population across all samples analyzed as indicated. Bar graphs of gene expression analysis represent mean +/- SEM or +/- SD as indicated. Two-tailed students unpaired t-test and Fisher's exact test (Excel, GraphPad software) were used to determine statistical significance in gene expression analysis data and single cell functional assays respectively. The statistical significance of the P-value was defined as follows for all P-value comparisons made: P>0.05 - not significant, P =0.01-0.05 - significant (*), P= 0.001-0.01 - very significant (**), P<0.001 - extremely significant (***)

Wilcoxon rank sum test was done using R. Kruskal-Wallis test, stratified by group was used to define significant differences between LMPP, MLP and GMP in the single cell functional assay in SGF15/2 condition and gave the following p-values: LMPP - 5x10^{-6}, MLP - 0.1725, GMP - 0.7395. Wilcoxon rank sum test confirmed that there was no outlier among single cell LMPPs coming from different CB donors. Prism software was used to plot the gene expression analysis and single cell in vitro data. LDA plots were generated using R and the lines represent the estimates calculated using ELDA software.

**Data availability**

Bulk RNA sequencing data have been deposited in Arrayexpress (https://www.ebi.ac.uk/arrayexpress/) with accession number E-MTAB-5456. Single cell RNA sequencing data accession number: GSE100618.

**References**


Figure 1

<table>
<thead>
<tr>
<th>Population</th>
<th>Immunophenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSC</td>
<td>Lin CD34+CD38+CD90+CD45RA+CD10-</td>
</tr>
<tr>
<td>MPP</td>
<td>Lin CD34+CD38+CD90+CD45RA+CD10-</td>
</tr>
<tr>
<td>LMPP</td>
<td>Lin CD34+CD38+CD90+CD45RA+CD10-</td>
</tr>
<tr>
<td>MLP</td>
<td>Lin CD34+CD38+CD90+CD45RA+CD10-</td>
</tr>
<tr>
<td>B/NK</td>
<td>Lin CD34+CD38+CD90+CD45RA+CD10-</td>
</tr>
<tr>
<td>GMP</td>
<td>Lin CD34+CD38+CD123+CD45RA+CD10</td>
</tr>
<tr>
<td>CMP</td>
<td>Lin CD34+CD38+CD123+CD45RA+CD10</td>
</tr>
<tr>
<td>MEP</td>
<td>Lin CD34+CD38+CD123+CD45RA+CD10</td>
</tr>
</tbody>
</table>

Table: Cord blood vs Bone marrow

<table>
<thead>
<tr>
<th>Population</th>
<th>Cord blood</th>
<th>Bone marrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD34+CD38-</td>
<td>15.4 ± 5.9</td>
<td>16.1 ± 10.3</td>
</tr>
<tr>
<td>CD38+CD10-</td>
<td>12.8 ± 5.2</td>
<td>15.8 ± 10.4</td>
</tr>
<tr>
<td>HSC</td>
<td>5.3 ± 2.7</td>
<td>11.4 ± 8.5</td>
</tr>
<tr>
<td>MPP</td>
<td>4.0 ± 2.3</td>
<td>3.7 ± 3.4</td>
</tr>
<tr>
<td>MLP</td>
<td>1.4 ± 1.0</td>
<td>0.2 ± 0.2</td>
</tr>
<tr>
<td>CD34+CD38-</td>
<td>84.6 ± 5.9</td>
<td>83.9 ± 10.3</td>
</tr>
<tr>
<td>CD38+CD10-</td>
<td>78.7 ± 7.0</td>
<td>78.4 ± 8.5</td>
</tr>
<tr>
<td>CMP</td>
<td>37.7 ± 11.3</td>
<td>34.7 ± 6.9</td>
</tr>
<tr>
<td>GMP</td>
<td>16.8 ± 6.1</td>
<td>17.8 ± 6.1</td>
</tr>
<tr>
<td>MEP</td>
<td>13.9 ± 7.5</td>
<td>14.6 ± 4.9</td>
</tr>
<tr>
<td>B/NK</td>
<td>2.0 ± 1.7</td>
<td>5.1 ± 7.5</td>
</tr>
</tbody>
</table>

Figure 1a: Cord blood populations with immunophenotypes.

Figure 1b: Log-log plots showing the distribution of cells based on FSC-A and CD45RA.

Figure 1c: Live cell analysis with FSC-W and Lin plots.

Figure 1d: Distribution of cells in the Cord blood with FSC-A and Lin plots.

Figure 1e: Distribution of cells in the Bone marrow with FSC-A and Lin plots.
Figure 3

**Figure 3**

**a** FACS sorting for progenitors

- CD34+ CB
- MS-5 + SCF, G-CSF, FLT3L, IL15, IL2, DuP-697 (SGF15/2 condition)

**b** Frequency of plated wells

<table>
<thead>
<tr>
<th>Type</th>
<th>LMPP</th>
<th>MLP</th>
<th>GMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>2</td>
<td>11</td>
<td>22</td>
</tr>
<tr>
<td>NK</td>
<td>3</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>M</td>
<td>5</td>
<td>194</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>394</td>
<td>4</td>
</tr>
<tr>
<td>B_NK</td>
<td>6</td>
<td>38</td>
<td>31</td>
</tr>
<tr>
<td>M_G</td>
<td>16</td>
<td>ND</td>
<td>7</td>
</tr>
<tr>
<td>B_M</td>
<td>10</td>
<td>392</td>
<td>32</td>
</tr>
<tr>
<td>B_G</td>
<td>13</td>
<td>789</td>
<td>38</td>
</tr>
<tr>
<td>M_NK</td>
<td>9</td>
<td>392</td>
<td>11</td>
</tr>
<tr>
<td>NK_G</td>
<td>15</td>
<td>ND</td>
<td>17</td>
</tr>
<tr>
<td>B_NK_M</td>
<td>16</td>
<td>392</td>
<td>44</td>
</tr>
<tr>
<td>B_NK_G</td>
<td>18</td>
<td>ND</td>
<td>49</td>
</tr>
<tr>
<td>B_M_G</td>
<td>22</td>
<td>ND</td>
<td>43</td>
</tr>
<tr>
<td>NK_M_G</td>
<td>23</td>
<td>ND</td>
<td>20</td>
</tr>
<tr>
<td>B_NK_M_G</td>
<td>29</td>
<td>ND</td>
<td>56</td>
</tr>
</tbody>
</table>

**c** Percentage of positive plated wells (LMPP)

- **dii** Percentage myeloid potential
  - **Total**
  - **Lyphoid**

**e** Uni-lineage

**f** Bi-lineage

**g** Multi-lineage

**h** Potentiality

**i** Potentiality

**j** Potentiality

**k** Ly outputs

**o** Potentials

<table>
<thead>
<tr>
<th>Potentials</th>
<th>LMPP</th>
<th>MLP</th>
<th>GMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lymphoid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B, NK, T, B+NK, NK+T, B+NK+T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myeloid</td>
<td>M, G, M+G</td>
<td>M, G</td>
<td>M, G, M+G</td>
</tr>
<tr>
<td>B+NK+G, NK+M+G, B+NK+G+T, B+M+G+T, B+NK+M+G+T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NK+G, B+M+G, B+NK+M, B+NK+G, B+NK+M+G</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NK+M, B+M+G, B+NK+M, B+NK+G, B+NK+M+G</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total no.**

|        | 24 | 8  | 14 |
Figure 4

(a) Subcutaneous transplantation of human MSC leads to subcutaneous humanized BM niche development. Direct intr ossicle transplantation of NSG results in NSG subcutaneous transplantation of human MSC.

(b) % human chimerism (per 1000 transplanted cells)

<table>
<thead>
<tr>
<th>LMPP</th>
<th>MLP</th>
<th>GMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>1.5</td>
<td>3.0</td>
<td>4.5</td>
</tr>
<tr>
<td>6.0</td>
<td>7.5</td>
<td>9.0</td>
</tr>
</tbody>
</table>

(c) % within human cells (CD45+HLA-ABC)

(d) 2 weeks post transplantation

LMPP

- CD15+
- CD15+/CD14+
- CD14+

MLP

- CD15
- CD15+/CD14+
- CD14+

GMP

- CD15
- CD15+/CD14+
- CD14+

(e) Flow cytometry plots for CD15+ CD15+/CD14+ CD14+ cells.
Figure 5

(a) All genes bs=1000

(b) 10000 ANOVA Genes

(c) 3000 ANOVA Genes

(d) 300 Variant Genes

(e) PC1 (13.44%) vs PC2 (15.34%)

(f) PC1 (16.66%) vs PC2 (12.42%)

(g) PC1 (25.83%) vs PC2 (13.07%)

(h) PC1 (30.00%) vs PC2 (15.34%)

(i) GMP Signature

(ii) MLP Signature

(iii) LMPP Signature

All genes labeled with different colors:
- HSC
- MPP
- LMPP
- MLP
- CMP
- GMP
- MEP

Genes table with number of DEG:
- TFE3
- ZBTB7A
- IL7R
- FAM129C
- MYCT1
- HHEX
- EGR1
- AIF1
- ITGAL
- ZMYND8
- ALOX15B
- IRF8
- MAML2
- CXXC5
- HLA-B
- MPO
- EGR1
- PRTN3
- F13A1
- RUNX3
- HBB
- MYBL2
- CDC45
- AURKB
- UBE2C
- PC1
- PC2
- PC3
- PC4
- PC5
- PC6

Genes with color coding:
- Red
- Blue
- Green
- Orange

Genes with specific values:
- 1097
- 178
- 630
- 256
- 1152
Figure 6

a

Cluster 1
Cluster 2
Cluster 3

Cell type: MLP LMPP GMP

Log-normalised expression

Cell type: MLP LMPP GMP

b

Cell type

Cluster 1
Cluster 2
Cluster 3

NUDT19
MME
HEMGN
AC3
C1D9
CLN9
AfrAP2
SPEC21
RXFP1
SPRY2
KIAA0487
LTB
AMICA1
SLC2A5
ABCA1
JCHAIN
SCN3A
EPRS
LPA6
USP9X
DERA
PROSC
CTSO
BCDKHB
EMC3
ZDHHC6
ZNF136
SLC3A7
DLG1
ZNF674–AS1
ARS2
ERS11
ELAVL4
BFC08
SERPINE1
SHOC2
NOLC1
CXR1
STRADB
TMPO
PEL1
VMA21
MCM5
CTSG
PRTN3
ELANE
MPO
ZFPAN5
TFRC
CERKL
HSP90
VIM
CPA3
EFC3A

PC1
PC2

Cluster 1
Cluster 2
Cluster 3

MLP LMPP GMP

PC1
PC2
Figure 7

**a** CD10 level

**b** LMPP

**c** LMPP

**d** CD45RA level

**e** LMPP

**f** Uni-lineage

**g** Bi-lineage

**h** Multi-lineage

**i** Potentiality

**j** Potentiality

**k** GMP

**l** GMP

**m** CD38 (44% CD38+)

**n** CD38 (15% CD38)
Figure 8

HSC → MPP

Bi- and multi-lineage lympho-myeloid progenitors

Bi- and tri-lineage progenitors

Uni-lineage progenitors

LMP

Lymphopoiesis

Myelopoiesis

Erythro-megakaryopoiesis

Bi- and multi-lineage lympho-myeloid progenitors

Bi- and tri-lineage progenitors

Uni-lineage progenitors

HSC

MPP

LMP

T-cell

B-cell

NK

Mono

Gr

BFU

E

MkP

MEP

T cell

B cell

NK cell

Monocyte

Granulocyte