LETTER

Musashi–2 attenuates AHR signalling to expand human haematopoietic stem cells

Stefan Rentas[1], Nicholas T. Holzapfel[1]*, Muluken S. Belew[1], Gabriel A. Pratt[2,3]*, Veronique Voisin[4], Brian T. Wilhelm[5], Gary D. Bader[4], Gene W. Yeo[2,3,6] & Kristin J. Hope[1]

Umbilical cord blood-derived haematopoietic stem cells (HSCs) are essential for many life-saving regenerative therapies. However, despite their advantages for transplantation, their clinical use is restricted because HSCs in cord blood are found only in small numbers[1]. Small molecules that enhance haematopoietic stem and progenitor cell (HSPC) expansion in culture have been identified[2,3], but in many cases their mechanisms of action or the nature of the pathways they impinge on are poorly understood. A greater understanding of the molecular circuitry that underpins the self-renewal of human HSCs will facilitate the development of targeted strategies that expand HSCs for regenerative therapies. Whereas transcription factor networks have been shown to influence the self-renewal and lineage decisions of human HSCs[4,5], the post-transcriptional mechanisms that guide HSC fate have not been closely investigated. Here we show that overexpression of the RNA-binding protein Musashi-2 (MSI2) induces multiple pro-self-renewal phenotypes, including a 17-fold increase in short-term repopulating cells and a net 23-fold ex vivo expansion of long-term repopulating HSCs. By performing a global analysis of MSI2–RNA interactions, we show that MSI2 directly attenuates aryl hydrocarbon receptor (AHR) signalling through post-transcriptional downregulation of canonical AHR pathway components in cord blood HSPCs. Our study gives mechanistic insight into RNA networks controlled by RNA-binding proteins that underlie self-renewal and provides evidence that manipulating such networks ex vivo can enhance the regenerative potential of human HSCs.

Control of translation by RNA-binding proteins in human HSCs and its potential to regulate HSC self-renewal remain underexplored. MSI2 is known to regulate mouse HSCs[6–8] and has been predicted to influence mRNA translation[9], so we investigated the role of MSI2 in post-transcriptionally controlling self-renewal of human HSPCs. The expression of *MSI2* mRNA was elevated in primitive cord blood HSPCs and decreased during differentiation, whereas the *MSI2* paralogue, *MSI1*, was not expressed (Extended Data Fig. 1a–f). Lentiviral overexpression of MSI2 resulted in a 1.5-fold increase in colony-forming units (CFU) relative to control cells, principally due to a 3.7-fold increase in the most primitive CFU-granulocyte erythrocyte monocyte megakaryocyte (GEMM) colony type (Extended Data Fig. 2a and Fig. 1a). Remarkably, 100% of MSI2-overexpressing CFU-GEMMs generated secondary colonies compared to only 40% of control CFU-GEMMs. In addition, MSI2 overexpression yielded three times as many colonies per re-seeded CFU-GEMM (Fig. 1b, c and Extended Data Fig. 2b). During *in vitro* culture, MSI2-overexpressing cells were 2.3- and 6-fold more abundant than control cells at the 7- and 21-day time points, respectively (Extended Data Fig. 2c, d). Moreover, after 7 days in culture, MSI2-overexpressing cells showed a cumulative 9.3-fold increase in colony-forming cells in the absence of changes in cell cycling or

death (Extended Data Fig. 2e–h). Together, our data demonstrate that enforced expression of MSI2 has potent self-renewal-inducing effects on early progenitors and promotes their *in vitro* expansion.

Short-term repopulating cells (STRCs) produce a transient multi-lineage graft in non-obese diabetic (NOD)/SCID *Il2rg*[null] (NSG) mice[10], and in patients these cells reconstitute granulocytes and platelets that are essential for preventing post-transplantation infection and bleeding[1]. MSI2-overexpressing STRCs yielded 1.8-fold more primitive CD34+ cells post-infection and a 17-fold increase in functional STRCs relative to control STRCs, as determined by limiting dilution analysis (LDA) of human chimaerism in mice 3 weeks after transplantation (Fig. 1d–f and Extended Data Fig. 3a, b). Furthermore, at a protracted engraftment
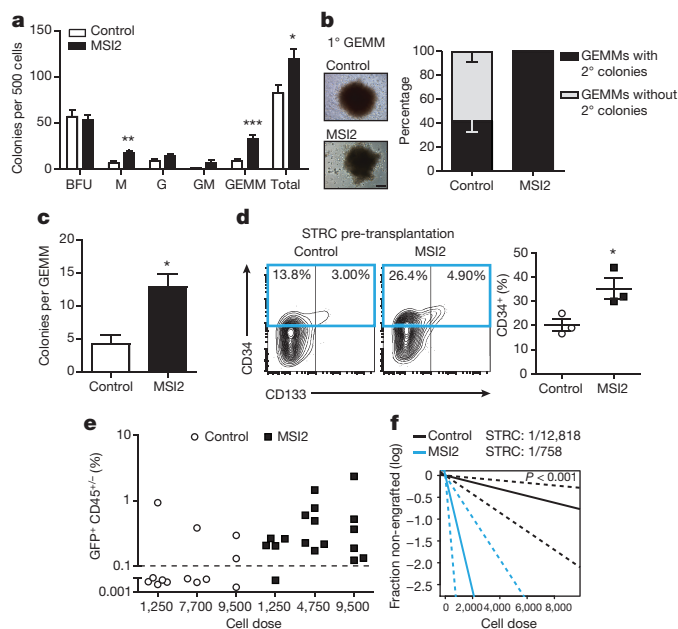

Figure 1 | MSI2 overexpression enhances in vitro cord blood progenitor activity and increases the number of STRCs. a, CFU output from transduced Lin− cord blood (n = 9 control and 10 MSI2-overexpressing (MSI2) cultures from 5 experiments). **b**, CFU-GEMM secondary CFU replating potential (n = 24 control and 30 MSI2-overexpressing cultures from 2 experiments) and images of primary GEMMs (scale bar, 200 μm). **c**, Number of secondary colonies per replated CFU-GEMM from **b**. **d**, CD34 expression in STRCs before transplantation (n = 3 experiments). **e**, Human chimaerism at 3 weeks in mice transplanted with varying doses of transduced STRCs. Dashed line indicates engraftment cut-off (n = 3 experiments). **f**, STRC frequency determined by LDA from **e**. Dashed lines indicate 95% confidence intervals. Data shown as mean ± s.e.m. Unpaired *t*-test, *P < 0.05; **P < 0.01; ***P < 0.001.

[1]Department of Biochemistry and Biomedical Sciences, Stem Cell and Cancer Research Institute, McMaster University, Hamilton, Ontario L8S 4K1, Canada. [2]Department of Cellular and Molecular Medicine, Institute for Genomic Medicine, University of California, San Diego, La Jolla, California 92037, USA. [3]Bioinformatics Graduate Program, University of California, San Diego, La Jolla, California 92037, USA. [4]The Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada. [5]Institute for Research in Immunology and Cancer, University of Montreal, Montreal, Quebec H3C 3J7, Canada. [6]Department of Physiology, National University of Singapore and Molecular Engineering Laboratory, A*STAR, Singapore 138632, Singapore.
*These authors contributed equally to this work.

readout time of 6.5 weeks at non-limiting transplant doses, 100% of mice transplanted with MSI2-overexpressing STRCs were engrafted compared to only 50% of mice transplanted with control STRCs, indicating that MSI2 overexpression extended the duration of STRC-mediated engraftment (Extended Data Fig. 3c).

We next explored the effect of short hairpin (sh)RNA-induced MSI2 knockdown on HSPC function. MSI2 knockdown did not alter the clonogenic potential of HSPCs but did decrease CFU replating threefold (Extended Data Fig. 4a–c). In more primitive culture-initiating cells, MSI2 knockdown significantly decreased cell numbers over culture (Extended Data Fig. 4d, e) independent of increased death or altered cell cycling (data not shown). Upon transplantation, engrafted MSI2 knockdown GFP+ cells showed no evidence of lineage skewing, but the frequency of cells was markedly reduced relative to the percentage of GFP+ cells initially transplanted (Extended Data Fig. 4f–h). Combined, our in vitro and in vivo data show that MSI2 knockdown reduces self-renewal in early progenitors and HSCs.

To characterize the earliest transcriptional changes induced by modulating MSI2 expression, we performed RNA sequencing (RNA-seq) on CD34+ MSI2-overexpressing and knockdown cells immediately after transduction (Supplementary Tables 1 and 2). MSI2 overexpression-induced transcriptional changes showed an inverse correlation with those induced by MSI2 knockdown, suggesting that overexpression and knockdown had opposite effects (Extended Data Fig. 5a). When compared to transcriptome data from 38 human haematopoietic cell subpopulations[4], genes that were significantly upregulated by MSI2 overexpression and downregulated upon MSI2 knockdown were exclusively enriched in those highly expressed in HSCs and other primitive CD34+ populations (Extended Data Fig. 5b).

As MSI2 overexpression conferred an HSC gene expression program, we hypothesized that it could facilitate HSC expansion ex vivo. MSI2 overexpression induced a fourfold increase in CD34+CD133+ phenotypic HSCs relative to control cells after 7 days of culture (Fig. 2a). We next performed an LDA to define functional HSC frequency before (day 3 post-transduction, D3) and after 7 days of ex vivo culture (day 10, D10; Extended Data Fig. 6a). Mice transplanted with HSCs on D3 displayed no altered engraftment as a result of MSI2 overexpression; however, recipients of MSI2-overexpressing D10-expanded cells displayed multiple phenotypes of enhanced reconstitution relative to recipients of control cells, including a twofold increase in bone marrow GFP+ levels without changes to lineage output, an increase in the proportion of GFP+ cells within the human graft relative to pre-transplant D10 levels, an increase in GFP mean fluorescence intensity and enrichment of CD34 expression in GFPhigh cells (Fig. 2b, c and Extended Data Fig. 6b–h). As the lentiviral construct design ensures that levels of GFP mirror those of MSI2, these findings indicate that high levels of MSI2 impart enhanced competitiveness and are conducive to in vivo HSPC activity. Importantly, D10 MSI2-overexpressing cultures contained more CD34+CD133+ cells before transplantation than did control cultures (Extended Data Fig. 6i) and, accordingly, the HSC frequency in D10 MSI2-overexpressing cultures was increased twofold relative to that in D3 MSI2-overexpressing cultures. By contrast, control cultures displayed a threefold decrease in HSC frequency. These results demonstrate that MSI2 overexpression ex vivo facilitated a net sixfold increase in HSC frequency relative to control cultures (Fig. 2g, h and Supplementary Tables 3, 4).

Secondary LDA transplants were performed to explore fully the effects of MSI2 overexpression and culturing on self-renewal and long-term HSCs (LT-HSCs). Robust engraftment with MSI2-overexpressing cells did not induce altered myelo-lymphopoiesis or leukaemic development (Fig. 2e). Secondary LDA measurements revealed that the percentage of GFP+ cells in the bone marrow was 4.6-fold higher after transplantation of MSI2-overexpressing cells than after control transplantation, and LT-HSC frequency was 3.5-fold higher (Fig. 2d, f and Supplementary Table 5). The increase in LT-HSC frequency corresponds to MSI2-overexpressing GFP+ HSCs having expanded in
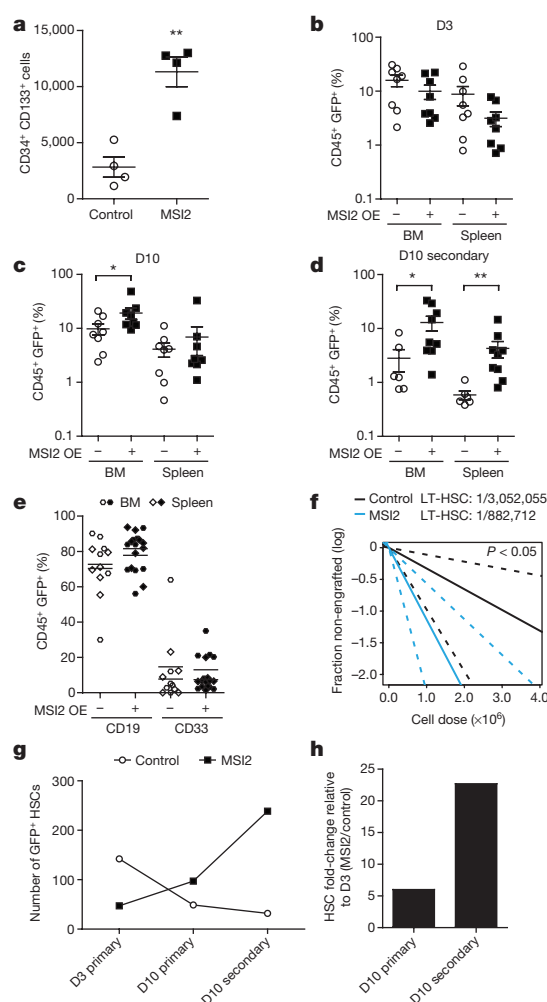

Figure 2 | MSI2 overexpression expands LT-HSCs in ex vivo culture.
a, Transduced CD34+CD133+ cells after 1 week of culture (n = 4 experiments, unpaired t-test). b–d, CD45+GFP+ engraftment in mice that received the highest two cell doses at D3 and D10 (n = 8 mice for both control and MSI2-overexpressing (MSI2 OE) cells) and the highest three cell doses of D10 secondary cells (n = 6 mice for control and 9 for MSI2-overexpressing cells, Mann–Whitney test). BM, bone marrow. e, Myelo-lymphopoiesis in mice that received D10 secondary cells. f, Multi-lineage LT-HSC frequency in bone marrow cells from mice that received D10 primary cells. Dashed lines indicate 95% confidence intervals. g, Numbers of GFP+ HSCs as evaluated by LDA. h, Cumulative fold change in MSI2-overexpressing HSCs. Data shown as mean ± s.e.m. *P < 0.05; **P < 0.01.

primary mice 2.4-fold over input as compared to a 1.5-fold decrease for control HSCs (Fig. 2g). The level of in vivo expansion induced by MSI2 overexpression reflects the behaviour of uncultured HSCs, which undergo similarly controlled expansion during passage in mice[3,11,12]. Finally, when accounting for the total change in GFP+ HSCs upon ex vivo culture, MSI2 overexpression induced a cumulative 23-fold expansion of secondary LT-HSCs relative to control (Fig. 2g, h), indicating that elevated MSI2 expression provides a considerable self-renewal advantage to functional HSCs during ex vivo culture.

To gain mechanistic insight into this process, we examined genes that were differentially expressed in MSI2-overexpressing cells and found that the gene encoding cytochrome P450 1B1 oxidase (CYP1B1), an effector of AHR signalling[13], was among the most repressed (Supplementary Table 1). Pathway analysis revealed that many predicted targets of AHR were enriched in the gene sets that were downregulated by MSI2 overexpression (Fig. 3a) and upregulated by MSI2 knockdown (Extended Data Fig. 7a, b). Binding of
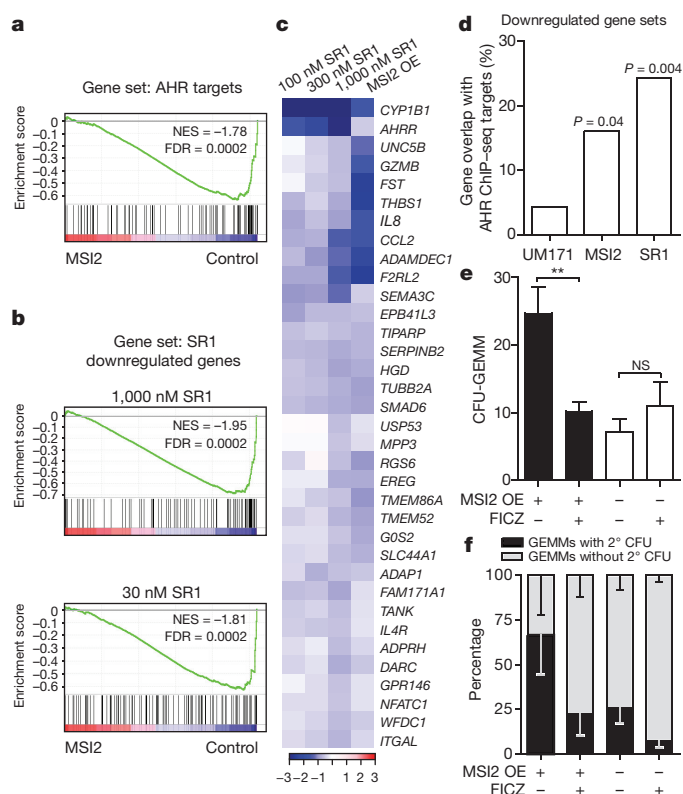
Figure 3 | MSI2 overexpression in human HSPCs attenuates AHR signalling. a, Predicted AHR targets compared by GSEA to genes downregulated by MSI2 overexpression. **b**, GSEA of gene sets downregulated by SR1 and MSI2 overexpression. **c**, log fold-change in expression of GSEA leading edge genes downregulated by MSI2 overexpression or SR1. **d**, Percentage of gene overlap between gene sets downregulated by UM171, SR1 or MSI2 overexpression and AHR targets identified by ChIP–seq. **e**, Number of CFU-GEMMs generated from transduced cells grown in CFU medium containing FICZ or dimethylsulfoxide (DMSO) ($n = 3$ experiments). **f**, CFU-GEMMs from **e** replated into CFU assays containing FICZ or DMSO ($n = 30$ control and 29 MSI2 overexpressing CFU-GEMMs per treatment). Data are presented as mean ± s.e.m. Unpaired t-test, **$P < 0.01$. NS, not significant.

the nuclear receptor transcription factor AHR to the small molecule StemRegenin 1 (SR1) inhibits AHR target gene activation and leads to the expansion of human HSPCs in culture[2]. Gene set enrichment analysis (GSEA) showed that genes that were downregulated by MSI2 overexpression significantly matched those that were downregulated by SR1 in an SR1 dose-dependent manner (Fig. 3b, c), whereas MSI2 knockdown induced the opposite expression profile (Extended Data Fig. 7c, d). We next examined the overlap between genes that were downregulated by MSI2 overexpression and AHR targets identified by chromatin-immunoprecipitation followed by sequencing (ChIP–seq)[14]. This comparison was extended to genes downregulated upon treatment with UM171, which expands HSPCs independently of AHR[3]. Direct transcriptional targets of AHR were enriched by 3.8- and 5.6-fold in the gene sets that were downregulated by MSI2 overexpression and SR1, respectively, compared to UM171; this overrepresentation was maintained for predicted AHR targets and suggests that MSI2 overexpression expands HSPCs by attenuating AHR signalling (Fig. 3d and Extended Data Fig. 7e). Furthermore, SR1 treatment increased the percentage of CD34+ cells eightfold in control cultures but only fourfold in MSI2-overexpressing cultures (Extended Data Fig. 8a, b), a finding that suggests that SR1 and MSI2 overexpression act redundantly on HSPCs via the same pathway.

To elucidate further the connection between MSI2 and AHR, MSI2-overexpressing and control cultures were treated with the AHR agonist 6-formylindolo(3,2-b)carbazole (FICZ).

Treatment of MSI2-overexpressing cells with FICZ induced canonical AHR targets, showing that these cells remain competent for AHR activation (Extended Data Fig. 8c). FICZ induced a marked reversal of the MSI2 overexpression-mediated increases in primary CFU-GEMMs and their replating capacities (Fig. 3e, f). Furthermore, FICZ-treated MSI2-overexpressing cultures displayed greater losses of phenotypic HSPCs compared to treated controls, which showed no change (Extended Data Fig. 8d, e). Together, these results show that agonist-induced restoration of AHR activity reduces the self-renewal-promoting effects of MSI2 overexpression and strongly supports the idea that MSI2 overexpression promotes HSPC expansion through downregulation of AHR signalling.

To identify key RNA targets that underlie MSI2 function, we analysed global MSI2 protein–RNA interactions using cross-linking immunoprecipitation followed by sequencing (CLIP–seq)[15] (Extended Data Fig. 9a, b). Replicates were highly correlated via gene RPKMs (reads per kilobase of transcript per million mapped reads) and 5,552 protein-coding genes were bound in both replicates (Extended Data Fig. 9c and Fig. 4a, b). Within the top 40% of reproducible clusters, MSI2 bound predominantly to the 3′ untranslated regions (3′UTRs) of mature mRNAs (Fig. 4c). Importantly, 9% of annotated protein-coding gene mRNAs were reproducible MSI2 targets, compared to 0.2% of long non-coding RNAs (Extended Data Fig. 9d), suggesting that MSI2 controls the stability or translation of coding mRNAs. Motif analysis identified a consensus pentamer (U/G)UAGU resembling the known mouse Msi1-binding sequence[9,16] within binding sites in all genic regions; additionally, MSI2-binding sites were generally significantly more conserved than background and tended to occur after the stop codon (Fig. 4d and Extended Data Fig. 9e–h). The presence of MSI2 binding sites within Msi1 targets[16] across species indicates that Musashi proteins may bind the same genes through 3′UTR-embedded motifs (Extended Data Fig. 9i). Finally, target gene ontology analysis revealed 186 biological processes categories (Supplementary Table 6), among the most significant of which were electron transport, oestrogen receptor signalling regulation and metabolism of small molecules, all processes known to be transcriptionally influenced by AHR signalling[17].

Among the top 2% of enriched CLIP–seq targets (Supplementary Table 7) were the 3′UTRs of the genes for two AHR pathway components: heat shock protein 90 (HSP90) and CYP1B1. Each exhibited multiple MSI2-binding motifs correlating with overlapping clusters of CLIP–seq reads (Fig. 4e and Extended Data Fig. 10a). To investigate the ability of MSI2 to post-transcriptionally regulate these genes during HSPC expansion, we looked for instances of uncoupled transcript and protein expression. HSP90 displayed uncoupling of transcript (1.6-fold up) and protein (1.6-fold down) expression early in culture, but after 7 days showed further upregulated transcript expression (2.5-fold) and variable protein levels (Fig. 4f and Extended Data Fig. 10b). As AHR–HSP90 binding is essential for ligand-dependent transcriptional activity[13], downregulation of HSP90 protein at the outset of HSPC culture would be expected to reduce latent AHR complex formation and attenuate AHR signalling (Fig. 3a). Indeed, CYP1B1 transcript and protein expression displayed twofold reductions early in culture, consistent with decreased AHR pathway activity; however, at day 7, CYP1B1 transcripts were upregulated 1.7-fold and uncoupled from protein expression, which was downregulated twofold (Fig. 4g and Extended Data Fig. 10c). To test whether MSI2 directly mediates post-transcriptional repression of these targets, the 3′UTRs of *CYP1B1* and *HSP90* were coupled to luciferase. MSI2 overexpression induced significant reductions in luciferase signal from both reporters, and this effect was mitigated when the core CLIP–seq-identified UAG motifs were mutated (Extended Data Fig. 10d, e). As MSI2 overexpression-mediated post-transcriptional downregulation of the AHR pathway converged on CYP1B1 protein repression throughout culture, we explored the effects on HSPCs of inhibiting CYP1B1 independently with (E)-2,3′,4,5′-tetramethoxystilbene (TMS). During culture, TMS
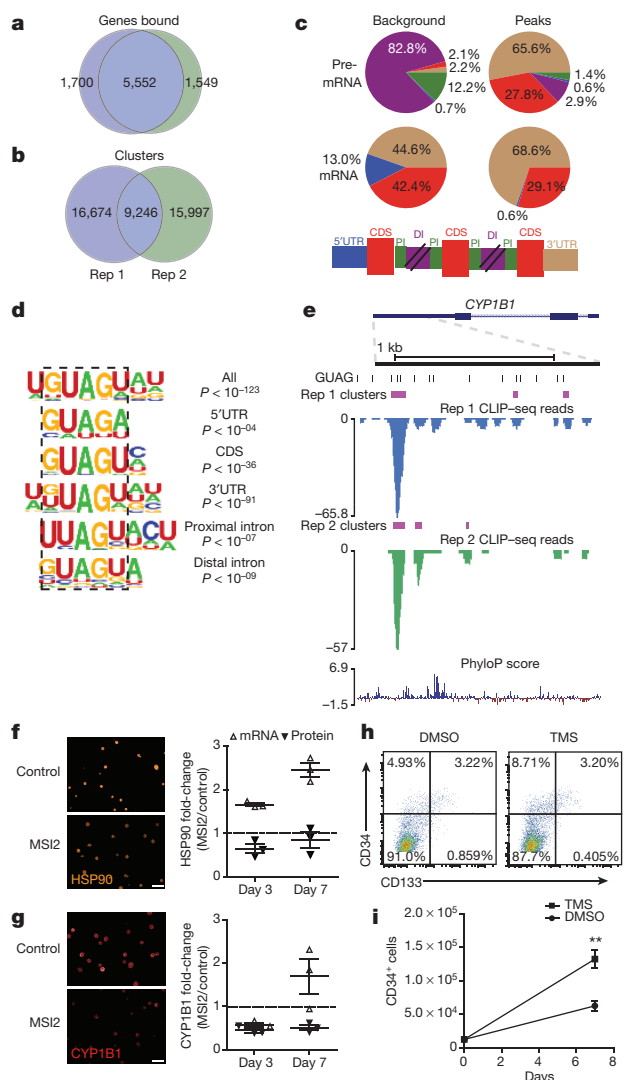
a Genes bound

1,700 | 5,552 | 1,549

b Clusters

16,674 | 9,246 | 15,997

Rep 1 | Rep 2

c Background | Peaks

Pre-mRNA 82.8% 2.1% 2.2% 12.2% 0.7% | 65.6% 1.4% 0.6% 2.9% 27.8%

13.0% mRNA 44.6% 42.4% | 68.6% 29.1% 0.6%

5'UTR CDS PI PI CDS PI PI DI CDS PI 3'UTR

d

UGUAGUAU All $P < 10^{-123}$
GUAGA 5'UTR $P < 10^{-04}$
GUAGUC CDS $P < 10^{-36}$
UGUAGUAU 3'UTR $P < 10^{-91}$
UUAGUACU Proximal intron $P < 10^{-07}$
GUAGUA Distal intron $P < 10^{-09}$

e CYP1B1

1 kb
GUAG
Rep 1 clusters
Rep 1 CLIP–seq reads 0 to −65.8
Rep 2 clusters
Rep 2 CLIP–seq reads 0 to −57
PhyloP score 6.9 to −1.5

f Control / MSI2 — HSP90

▲ mRNA ▼ Protein
HSP90 fold-change (MSI2/control)
Day 3 | Day 7

g Control / MSI2 — CYP1B1

CYP1B1 fold-change (MSI2/control)
Day 3 | Day 7

h DMSO | TMS
CD34 / CD133
DMSO: 4.93% 3.22% 91.0% 0.859%
TMS: 8.71% 3.20% 87.7% 0.405%

i
CD34+ cells
2.0 × 10^5, 1.5 × 10^5, 1.0 × 10^5, 5.0 × 10^4
TMS / DMSO
**
0 2 4 6 8 Days

Figure 4 | MSI2 overexpression post-transcriptionally downregulates AHR pathway components. a, Overlap between MSI2 target genes from separate CLIP–seq experiments. **b**, Statistically significant overlap ($P < 0.0001$, hypergeometric test) of clusters between the replicates. **c**, Percentage of CLIP–seq clusters in different genic regions. **d**, Consensus motifs within MSI2 clusters in different genic regions. *P* values presented for the top 40% of clusters. **e**, CLIP–seq reads (blue, replicate 1; green, replicate 2) and clusters (purple) mapped to the 3'UTR of *CYP1B1*. Matches to the GUAG motif are shown in black. **f, g**, Immunofluorescence for HSP90 and CYP1B1 3 days after transduction and summary of fold-changes in HSP90 and CYP1B1 protein and transcript levels with MSI2 overexpression at 3 and 7 days after transduction (scale bar, 20 μm; dotted line indicates no change; $n = 3$ experiments). **h**, HSPC marker expression by CD34+ cells treated with TMS for 10 days. **i**, Absolute CD34+ cell number with TMS ($n = 4$ experiments). Data are presented as mean ± s.e.m. Unpaired *t*-test, **$P < 0.01$.

increased the frequency and total numbers of CD34+ cells by 1.5-fold and 2-fold, respectively (Fig. 4h, i), phenocopying the effects of MSI2 overexpression. Finally, overexpression of both CYP1B1 lacking its 3'UTR and MSI2 decreased secondary CFU-GEMM replating efficiency (Extended Data Fig. 10f, g); this suggests that CYP1B1, while typically used to report AHR signalling, itself promotes HSPC differentiation.

Our work identifies MSI2 as an important mediator of human HSPC self-renewal and *ex vivo* expansion that acts by coordinating the post-transcriptional regulation of proteins belonging to a shared self-renewal regulatory pathway (Extended Data Fig. 10h). We envision that manipulation of the post-transcriptional circuitry controlled by RNA-binding proteins will provide a novel and powerful means by which to enhance the regenerative potential of not only human HSCs but also other stem-cell types.

1. Miller, P. H., Knapp, D. J. & Eaves, C. J. Heterogeneity in hematopoietic stem cell populations: implications for transplantation. *Curr. Opin. Hematol.* **20,** 257–264 (2013).
2. Boitano, A. E. *et al.* Aryl hydrocarbon receptor antagonists promote the expansion of human hematopoietic stem cells. *Science* **329,** 1345–1348 (2010).
3. Fares, I. *et al.* Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science* **345,** 1509–1512 (2014).
4. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144,** 296–309 (2011).
5. Laurenti, E. *et al.* The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nature Immunol.* **14,** 756–763 (2013).
6. Hope, K. J. *et al.* An RNAi screen identifies Msi2 and Prox1 as having opposite roles in the regulation of hematopoietic stem cell activity. *Cell Stem Cell* **7,** 101–113 (2010).
7. de Andrés-Aguayo, L. *et al.* Musashi 2 is a regulator of the HSC compartment identified by a retroviral insertion screen and knockout mice. *Blood* **118,** 554–564 (2011).
8. Park, S. M. *et al.* Musashi-2 controls cell fate, lineage bias, and TGF-β signaling in HSCs. *J. Exp. Med.* **211,** 71–87 (2014).
9. Ohyama, T. *et al.* Structure of Musashi1 in a complex with target RNA: the role of aromatic stacking interactions. *Nucleic Acids Res.* **40,** 3218–3231 (2012).
10. Glimm, H. *et al.* Previously undetected human hematopoietic cell populations with short-term repopulating activity selectively engraft NOD/SCID-β2 microglobulin-null mice. *J. Clin. Invest.* **107,** 199–206 (2001).
11. Cashman, J. D. & Eaves, C. J. Human growth factor-enhanced regeneration of transplantable human hematopoietic stem cells in nonobese diabetic/severe combined immunodeficient mice. *Blood* **93,** 481–487 (1999).
12. Holyoake, T. L., Nicolini, F. E. & Eaves, C. J. Functional differences between transplantable human hematopoietic stem cells from fetal liver, cord blood, and adult marrow. *Exp. Hematol.* **27,** 1418–1427 (1999).
13. Mimura, J. & Fujii-Kuriyama, Y. Functional role of AhR in the expression of toxic effects by TCDD. *Biochim. Biophys. Acta* **1619,** 263–268 (2003).
14. Lo, R. & Matthews, J. High-resolution genome-wide mapping of AHR and ARNT binding sites by ChIP-seq. *Toxicol. Sci.* **130,** 349–361 (2012).
15. Yeo, G. W. *et al.* An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Struct. Mol. Biol.* **16,** 130–137 (2009).
16. Katz, Y. *et al.* Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state. *Elife* **3,** e03915 (2014).
17. Tijet, N. *et al.* Aryl hydrocarbon receptor regulates distinct dioxin-dependent and dioxin-independent gene batteries. *Mol. Pharmacol.* **69,** 140–153 (2006).

Author Contributions S.R. designed and performed experiments, analysed data and wrote the manuscript. N.T.H. constructed CLIP–seq libraries. M.S.B. helped perform cord blood experiments. G.A.P. and G.W.Y advised on CLIP–seq library construction, performed CLIP–seq bioinformatic analyses and wrote the manuscript. B.T.W. performed RNA-seq analyses. V.V. and G.D.B performed RNA-seq bioinformatic analyses. K.J.H. conceived the project, supervised the study, analysed data, interpreted results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.W.Y. (geneyeo@ucsd.edu) or K.J.H. (kristin@mcmaster.ca).

METHODS

Mice. NOD/SCID $Il2rg^{null}$ mice (Jackson Laboratory) were bred and maintained in the Stem Cell Unit animal barrier facility at McMaster University. All procedures were approved by the Animal Research Ethics Board at McMaster University.

Isolation of primitive human haematopoietic cells and flow cytometry. All patient samples were obtained with informed consent and with the approval of local human subject research ethics boards at McMaster University. Human umbilical cord blood mononuclear cells were collected by centrifugation with Ficoll-Paque Plus (GE), followed by red blood cell lysis with ammonium chloride (StemCell Technologies). Cells were then incubated with a cocktail of lineage-specific antibodies (CD2, CD3, CD11b, CD11c, CD14, CD16, CD19, CD24, CD56, CD61, CD66b, and GlyA; StemCell Technologies) for negative selection of Lin− cells using an EasySep immunomagnetic column (StemCell Technologies). Live cells were discriminated on the basis of cell size, granularity and, as needed, absence of viability dye 7-AAD (BD Biosciences) uptake. All flow cytometry analysis was performed using a BD LSR II instrument (BD Biosciences). Data acquisition was conducted using BD FACSDiva software (BD Biosciences) and analysis was performed using FlowJo software (Tree Star).

HSPC sorting and qRT–PCR analysis. To quantify *MSI2* expression in human HSPCs, Lin− cord blood cells were stained with the appropriate antibody combinations to resolve HSC (CD34+ CD38− CD45RA− CD90+), MPP (CD34+ CD38− CD45RA− CD90−), CMP (CD34+ CD38+ CD71−) and EP (CD34+ CD38+ CD71+) fractions as similarly described previously[18,19] with all antibodies from BD Biosciences: CD45RA (HI100), CD90 (5E10), CD34 (581), CD38 (HB7) and CD71 (M-A712). Cell viability was assessed using the viability dye 7AAD (BD Biosciences). All cell subsets were isolated using a BD FACSAria II cell sorter (BD Biosciences) or a MoFlo XDP cell sorter (Beckman Coulter). HemaExplorer[20] analysis was used to confirm *MSI2* expression in human HSPCs and across the hierarchy. For all qRT–PCR determinations total cellular RNA was isolated with TRIzol LS reagent according to the manufacturer's instructions (Invitrogen) and cDNA was synthesized using the qScript cDNA Synthesis Kit (Quanta Biosciences). qRT–PCR was done in triplicate with PerfeCTa qPCR SuperMix Low ROX (Quanta Biosciences) with gene-specific probes (Universal Probe Library (UPL), Roche) and primers: MSI2 UPL-26, F-GGCAGCAAGAGGATCAGG, R-CCGTAGAGATCGGCGACA; HSP90 UPL-46, F-GGGCAACACCTCTACAAGGA, R-CTTGGGTCTGGGTTTCCTC; CYP1B1 UPL-20, F-ACGTACCGGCCACTATCACT, R-CTCGAGTCTGCAC ATCAGGA; GAPDH UPL-60, F-AGCCACATCGCTCAGACAC, R-GCCCAA TACGACCAAATCC; ACTB (UPL Set Reference Gene Assays, Roche). The mRNA content of samples compared by qRT–PCR was normalized based on the amplification of *GAPDH* or *ACTB*.

Lentivirus production and western blot validation. MSI2 shRNAs were designed with the Dharmacon algorithm (http://www.dharmacon.com). Predicted sequences were synthesized as complimentary oligonucleotides, annealed and cloned downstream of the H1 promoter of the modfied cppt-PGK-EGFP-IRES-PAC-WPRE lentiviral expression vector[18]. Sequences for the MSI2 targeting and control RFP targeting shRNAs were as follows: shMSI2, 5′-GAGAGATCCCACTACGAAA-3′; shRFP, 5′-GTGGGAGCGCGTGATGAAC-3′. Human *MSI2* cDNA (BC001526; Open Biosystems) was subcloned into the MA bi-directional lentiviral expression vector[21]. Human *CYP1B1* cDNA (BC012049; Open Biosystems) was cloned in to psMALB[22]. All lentiviruses were prepared by transient transfection of 293FT (Invitrogen) cells with pMD2.G and psPAX2 packaging plasmids (Addgene) to create VSV-G pseudotyped lentiviral particles. All viral preparations were titrated on HeLa cells before use on cord blood. Standard SDS–PAGE and western blotting procedures were performed to validate the effects of knockdown on transduced NB4 cells (DSMZ) and overexpression on 293FT cells. Immunoblotting was performed with anti-MSI2 rabbit monoclonal IgG (EP1305Y, Epitomics) and β-actin mouse monoclonal IgG (ACTBD11B7, Santa Cruz Biotechnology) antibodies. Secondary antibodies used were IRDye 680 goat anti-rabbit IgG and IRDye 800 goat anti-mouse IgG (LI-COR). 293FT and NB4 cell lines tested negative for mycoplasma. NB4 cells were authenticated by ATRA treatment before use.

Cord blood transduction. Cord blood transductions were conducted as described previously[18,23]. Briefly, thawed Lin− cord blood or flow-sorted Lin− CD34+ CD38− or Lin− CD34+ CD38+ cells were prestimulated for 8–12 h in StemSpan medium (StemCell Technologies) supplemented with growth factors interleukin 6 (IL-6; 20 ng ml−1, Peprotech), stem cell factor (SCF; 100 ng ml−1, R&D Systems), Flt3 ligand (FLT3-L; 100 ng ml−1, R&D Systems) and thrombopoietin (TPO; 20 ng ml−1, Peprotech). Lentivirus was then added in the same medium at a multiplicity of infection of 30–100 for 24 h. Cells were then given 2 days after transduction before use in *in vitro* or *in vivo* assays. For *in vitro* cord blood studies biological (experimental) replicates were performed with three independent cord blood samples.

Clonogenic progenitor assays. Human clonogenic progenitor cell assays were done in semi-solid methylcellulose medium (Methocult H4434; StemCell

Technologies) with flow-sorted GFP+ cells post transduction (500 cells per ml) or from day seven cultured transduced cells (12,000 cells per ml). Colony counts were carried out after 14 days of incubation. CFU-GEMMs can seed secondary colonies owing to their limited self-renewal potential[24]. Replating of MSI2-overexpressing and control CFU-GEMMs for secondary CFU analysis was performed by picking single CFU-GEMMs at day 14 and disassociating colonies by vortexing. Cells were spun and resuspended in fresh methocult, mixed with a blunt-ended needle and syringe, and then plated into single wells of a 24-well plate. Secondary CFU analysis for shMSI2- and shControl-expressing cells was performed by harvesting total colony growth from a single dish (as nearly equivalent numbers of CFU-GEMMs were present in each dish), resuspending cells in fresh methocult by mixing vigorously with a blunt-ended needle and syringe and then plating into replicate 35-mm tissue culture dishes. In both protocols, secondary colony counts were done following incubation for 10 days. For primary and secondary colony forming assays performed with the AHR agonist FICZ (Santa Cruz Biotechnology), 200 nM FICZ or 0.1% DMSO was added directly to H4434 methocult medium. Two-way ANOVA analysis was performed to compare secondary CFU output for MSI2-overexpressing or control conditions and FICZ treatment for MSI2-overexpressing or control conditions. Colonies were imaged with a Q-Colour3 digital camera (Olympus) mounted to an Olympus IX5 microscope with a 10× objective lens. Image-Pro Plus imaging software (Media Cybernetics) was used to acquire pictures and subsequent image processing was performed with ImageJ software (NIH).

Lin− cord blood and Lin− CD34+ suspension cultures. Transduced human Lin− cord blood cells were sorted for GFP expression and seeded at a density of 10^5 cells per ml in IMDM 10% FBS supplemented with human growth factors IL-6 (10 ng ml−1), SCF (50 ng ml−1), FLT3-L (50 ng ml−1), and TPO (20 ng ml−1) as previously described[25]. To generate growth curves, every seven days cells were counted, washed, and resuspended in fresh medium with growth factors at a density of 10^5 cells per ml. Cells from suspension cultures were also used in clonogenic progenitor, cell cycle and apoptosis assays. Experiments performed on transduced Lin− CD34+ cord blood cells used serum-free conditions as described in the cord blood transduction subsection of Methods. For *in vitro* cord blood studies, biological (experimental) replicates were performed with three independent cord blood samples.

Cell cycle and apoptosis assays. Cell cycle progression was monitored with the addition of BrdU to day 10 suspension cultures at a final concentration of 10 μM. After 3 h of incubation, cells were assayed with the BrdU Flow Kit (BD Biosciences) according to the manufacturer's protocol. Cell proliferation and quiescence were measured using Ki67 (BD Bioscience) and Hoechst 33342 (Sigma) on day 4 suspension cultures after fixing and permeabilizing cells with the Cytofix/Cytoperm kit (BD Biosciences). For apoptosis analysis, Annexin V (Invitrogen) and 7-AAD (BD Bioscience) staining of day 7 suspension cultures was performed according to the manufacturer's protocol.

Intracellular flow cytometry. Lin− cord blood cells were initially stained with anti-CD34 PE (581) and anit-CD38 APC (HB7) antibodies (BD Biosciences) then fixed with the Cytofix/Cytoperm kit (BD Biosciences) according to the manufacturer's instructions. Fixed and permeabilized cells were immunostained with anti-MSI2 rabbit monoclonal IgG antibody (EP1305Y, Abcam) and detected by Alexa-488 goat anti-rabbit IgG antibody (Invitrogen).

RNA-seq data processing. CD34+ cells were transduced with an MSI2-overexpression or MSI2-knockdown lentivirus along with their corresponding controls and sorted for GFP expression 3 days later. Transductions for MSI2 overexpression or knockdown were each performed on two independent cord blood samples. Total RNA from transduced cells ($>1 \times 10^5$) was isolated using TRIzol LS as recommended by the manufacturer (Invitrogen), and then further purified using RNeasy columns (Qiagen). Sample quality was assessed using Bioanalyzer RNA Nano chips (Agilent). Paired-end, barcoded RNA-seq sequencing libraries were then generated using the TruSeq RNA Sample Prep Kit (v2) (Illumina) following the manufacturer's protocols starting from 1 μg total RNA. The quality of library generation was then assessed using a Bioanalyzer platform (Agilent) and Illumina MiSeq-QC run was performed or quantified by qPCR using KAPA quantification kit (KAPA Biosystems). Sequencing was performed using an Illumina HiSeq2000 using TruSeq SBS v3 chemistry at the Institute for Research in Immunology and Cancer's Genomics Platform (University of Montreal) with cluster density targeted at 750,000 clusters per mm² and paired-end 2 × 100-bp read lengths. For each sample, 90–95 million reads were produced and mapped to the hg19 (GRCh37) human genome assembly using CASAVA (version 1.8). Read counts generated by CASAVA were processed in EdgeR (edgeR_3.12.0, R 3.2.2) using TMM normalization, paired design, and estimation of differential expression using a generalized linear model (glmFit). The false discovery rate (FDR) was calculated from the output *P* values using the Benjamini–Hochberg method. The fold change of logarithm of base 2 of TMM normalized data (logFC) was used to rank the data from top upregulated to top downregulated genes and FDR (0.05) was used to define

significantly differentially expressed genes. RNA-seq data have been deposited in NCBI's Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE70685.

GSEA and iRegulon AHR target prediction. iRegulon[26] was used to retrieve the top 100 AHR predicted targets with a minimal occurrence count threshold of 5. The data were analysed using GSEA[27] with ranked data as input with parameters set to 2,000 gene-set permutations.

GSEA and StemRegenin 1 (SR1) gene sets. The GEO dataset GSE28359, which contains Affymetrix Human Genome U133 Plus 2.0 Array gene expression data for CD34[+] cells treated with SR1 at 30 nM, 100 nM, 300 nM and 1,000 nM was used to obtain lists of genes differentially expressed in the treated samples compared to the control ones (0 nM)[2]. Data were background corrected using Robust Multi-Array Average (RMA) and quantile normalized using the *expresso()* function of the affy Bioconductor package (affy_1.38.1, R 3.0.1). Lists of genes were created from the 150 top upregulated and downregulated genes from the SR1-treated samples at each dose compared to the non-treated samples (0 nM). The data were analysed using GSEA with ranked data as input with parameters set to 2,000 gene-set permutations. The normalized enrichment score (NES) and false discovery rate (FDR) were calculated for each comparison.

Differentiation Map of Haematopoiesis (DMAP) population comparisons. The GEO data set GSE24759, which contains Affymetrix GeneChip HT-HG_U133A Early Access Array gene expression data for 38 distinct haematopoietic cell states[4], was compared to the MSI2 overexpression and knockdown data. GSE24759 data were background corrected using Robust Multi-Array Average (RMA), quantile normalized using the *expresso()* function of the affy Bioconductor package (affy_1.38.1, R 3.0.1), batch corrected using the *ComBat()* function of the sva package (sva_3.6.0) and scaled using the standard score. Bar graphs were created by calculating for significantly differentially expressed genes the number of scaled data that were above (>0) or below (<0) the mean for each population. Percentages indicating for how long the observed value (set of up- or downregulated genes) was better represented in that population than random values were calculated from 1,000 trials.

AHR ChIP–seq comparison with downregulated gene sets. A unique list of genes closest to AHR-bound regions previously identified from TCDD-treated MCF7 ChIP–seq data[14] was used to calculate the overlap with genes showing >1.5-fold downregulation in response to treatment with UM171 (35 nM) or SR1 (500 nM) relative to DMSO-treated samples[3] as well as with genes significantly downregulated in MSI2-overexpressing versus control treated samples (FDR < 0.05). The percentage of downregulated genes with AHR-bound regions was then plotted for each gene set. *P* values were generated with Fisher's exact test for comparisons between gene lists.

oPOSSUM analysis for promoter AHR binding sites in downregulated gene sets. AHR transcription factor binding sites in downregulated gene sets were identified with oPOSSUM-3[28]. Genes showing >1.5-fold downregulation in response to treatment with UM171 (35 nM) or SR1 (500 nM) relative to DMSO-treated samples[3] were used along with significantly downregulated genes (FDR < 0.05) with EdgeR-analysed MSI2-overexpressing versus control-treated samples. The three gene lists were uploaded into oPOSSUM-3 and the AHR:ARNT transcription factor binding site profile was used with the matrix score threshold set at 80% to analyse the region 1,500 bp upstream and 1,000 bp downstream of the transcription start site. The percentage of downregulated genes with AHR-binding sites in their promoters was then plotted for each gene set. Fisher's exact test was used to identify significant overrepresentation of AHR-binding sites in gene lists relative to background.

Analysis for human chimaerism. Eight- to 12-week-old male or female NSG mice were sublethally irradiated (315 cGy) one day before intrafemoral injection with transduced cells carried in IMDM 1% FBS at 25 μl per mouse. Injected mice were analysed for human haematopoietic engraftment 12–14 weeks after transplantation or at 3 and 6.5 weeks for STRC experiments. Mouse bones (femurs, tibiae and pelvis) and spleen were removed and bones were crushed with a mortar and pestle then filtered into single-cell suspensions. Bone marrow and spleen cells were blocked with mouse Fc block (BD Biosciences) and human IgG (Sigma) and then stained with fluorochrome-conjugated antibodies specific to human haematopoietic cells. For multilineage engraftment analysis, cells from mice were stained with CD45 (HI30) (Invitrogen), CD33 (P67.6), CD15 (HI98), CD14 (MφP9), CD19 (HIB19), CD235a/GlyA (GA-R2), CD41a (HIP8) and CD34 (581) (BD Biosciences).

HSC and STRC xenotransplantation. For MSI2 knockdown in HSCs, 5.0×10^4 and 2.5×10^4 sorted Lin[−] CD34[+] CD38[−] cells were used per short-hairpin transduction experiment, leading to transplantation of day zero equivalent cell doses of 10×10^3 and 6.25×10^3, respectively, per mouse. For STRC LDA transplantation experiments, 10^5 sorted CD34[+]CD38[+] cells were used per control or MSI2-overexpressing transduction. After assessing levels of gene transfer, day zero equivalent GFP[+] cell doses were calculated to perform the LDA. Recipients with greater than 0.1% GFP[+]CD45[+/−] cells were considered to be repopulated. For STRC experiments that read out extended engraftment at 6.5 weeks, 2×10^5 CD34[+] CD38[+] cells were used per overexpressing or control transduction to allow

non-limiting 5×10^4 day zero equivalent cell doses per mouse. For HSC expansion and LDA experiments, CD34[+]CD38[−] cells were sorted and transduced with MSI2-overexpressing or control vectors (50,000 cells per condition) for 3 days and then analysed for gene-transfer levels (% GFP[+/−]) and primitive cell marker expression (% CD34 and CD133). To ensure that equal numbers of GFP[+] cells were transplanted into both control and MSI2-overexpressing recipient mice, we added identically cultured GFP[−] cells to the MSI2 culture to match the % GFP[+] of the control culture (necessary owing to the differing efficiency of transduction). The adjusted MSI2-overexpressing culture was recounted and aliquoted (63,000 cells) to match the output of half of the control culture. Three day 0 equivalent GFP[+] cell doses (1,000, 300 and 62 cells) were then transplanted per mouse to perform the D3 primary LDA. A second aliquot of the adjusted MSI2-overexpressing culture was then taken and put into culture in parallel with the remaining half of the control culture to perform another LDA after 7 days of growth (10 days total growth, D10 primary LDA). Altogether, four cell doses were transplanted; when converted back to day 0 equivalents these equalled approximately 1,000, 250, 100, and 20 GFP[+] cells per mouse, respectively. Pooled bone marrow from six engrafted primary mice that received D10 cultured control or MSI2-overexpressing cells (from the two highest doses transplanted) was aliquoted into five cell doses of 15 million, 10 million, 6 million, 2 million and 1 million cells. The numbers of GFP[+] cells within primary mice was estimated from nucleated cell counts obtained from NSG femurs, tibias and pelvises and from Colvin *et al.*[29]. The actual numbers of GFP[+] cells used for determining numbers of GFP[+] HSCs and the number of mice transplanted for all LDA experiments is shown in Supplementary Tables 3–5. The cut-off for HSC engraftment was a demonstration of multilineage reconstitution that was set at bone marrow having >0.1% GFP[+] CD33[+] and >0.1% GFP[+] CD19[+] cells. HSC and STRC frequency was assessed using ELDA software[30]. For all mouse transplantation experiments, mice were age- (6–12 week) and sex-matched. All transplanted mice were included for analysis unless mice died from radiation sickness before the experimental endpoint. No randomization or blinding was performed for animal experiments. Approximately 3–6 mice were used per cell dose for each cord blood transduction and transplantation experiment.

UV CLIP–seq library preparation. CLIP–seq was performed as previously described[15]. Briefly, 25 million NB4 cells (a transformed human cell line of haematopoietic origin) were washed in PBS and UV-cross-linked at 400 mJ cm[−2] on ice. Cells were pelleted, lysed in wash buffer (PBS, 0.1% SDS, 0.5% Na-deoxycholate, 0.5% NP-40) and DNase-treated, and supernatants from lysates were collected for immunoprecipitation. MSI2 was immunoprecipated overnight using 5 μg of anti-MSI2 antibody (EP1305Y, Abcam) and Protein A Dynabeads (Invitrogen). Beads containing immunoprecipated RNA were washed twice with wash buffer, high-salt wash buffer (5× PBS, 0.1% SDS, 0.5% Na-Deoxycholate, 0.5% NP-40), and PNK buffer (50 mM Tris-Cl pH 7.4, 10 mM MgCl₂, 0.5% NP-40). Samples were then treated with 0.2 U MNase for 5 min at 37° with shaking to trim immunopreciptated RNA. MNase inactivation was then carried out with PNK + EGTA buffer (50 mM Tris-Cl pH 7.4, 20 mM EGTA, 0.5% NP-40). The sample was dephosphorylated using alkaline phosphatase (CIP, NEB) at 37° for 10 min followed by washing with PNK+EGTA, PNK buffer, and then 0.1 mg ml[−1] BSA in nuclease-free water. 3′RNA linker ligation was performed at 16° overnight with the following adaptor: 5′P-UGGAAUUCUCGGGUGCCAAGG-puromycin. Samples were then washed with PNK buffer, radiolabelled using P32-γ-ATP (Perkin Elmer), run on a 4–12% Bis-Tris gel and then transferred to a nitrocellulose membrane. The nitrocellulose membrane was developed via autoradiography and RNA–protein complexes 15–20 kDa above the molecular weight of MSI2 were extracted with proteinase K followed by RNA extraction with acid phenol-chloroform. A 5′RNA linker (5′HO-GUUCAGAGUUCUACAGUCCGACGAUC-OH) was ligated to the extracted RNA using T4 RNA ligase (Fermentas) for 2 h and the RNA was again purified using acid phenol-chloroform. Adaptor ligated RNA was re-suspended in nuclease-free water and reverse transcribed using Superscript III reverse transcriptase (Invitrogen). Twenty cycles of PCR were performed using NEB Phusion Polymerase using a 3′PCR primer that contained a unique Illumina barcode sequence. PCR products were run on an 8% TBE gel. Products ranging between 150 and 200 bp were extracted using the QIAquick gel extraction kit (Qiagen) and re-suspended in nuclease-free water. Two separate libraries were prepared and sent for single-end 50-bp Illumina sequencing at the Institute for Genomic Medicine at the University of California, San Diego. 47,098,127 reads from the first library passed quality filtering, of which 73.83% mapped uniquely to the human genome. 57,970,220 reads from the second library passed quality filtering, of which 69.53% mapped uniquely to the human genome. CLIP-data reproducibility was verified through high correlation between gene RPKMs and statistically significant overlaps in the clusters and genes within replicates. CLIP–seq data have been deposited in NCBI's GEO and are accessible through GEO Series accession number GSE69583.

CLIP–seq mapping and cluster identification. Before sequence alignment of CLIP–seq reads to the human genome was performed, sequencing reads from

libraries were trimmed of polyA tails, adapters, and low quality ends using Cutadapt with parameters–match-read-wildcards–times 2 -e 0 -O 5–quality-cutoff' 6 -m 18 -b TCGTATGCCGTCTTCTGCTTG -b ATCTCGTATGCCGTCTTCTGCTTG -b CGACAGGTTCAGAGTTCTACAGTCCGACGATC -b TGGAATTCTC GGGTGCCAAGG -b AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAA-b TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT TTTTTTTTTTTTTTTTTTTTT. Reads were then mapped against a database of repetitive elements derived from RepBase (version 18.05). Bowtie (version 1.0.0) with parameters -S -q -p 16 -e 100 -l 20 was used to align reads against an index generated from Repbase sequences[31]. Reads not mapped to Repbase sequences were aligned to the hg19 human genome (UCSC assembly) using STAR (version 2.3.0e)[32] with parameters–outSAMunmapped Within –outFilterMultimapNmax 1 –outFilterMultimapScoreRange 1. To identify clusters in the genome of significantly enriched CLIP–seq reads, reads that were PCR replicates were removed from each CLIP–seq library using a custom script of the same method as in ref. 33; otherwise, reads were kept at each nucleotide position when more than one read's 5′-end was mapped. Clusters were then assigned using the CLIPper software with parameters–bonferroni–superlocal–threshold-[34]. The ranked list of significant targets was calculated assuming a Poisson distribution, where the observed value is the number of reads in the cluster, and the background is the number of reads across the entire transcript and or across a window of 1000 bp ± the predicted cluster.

Gene annotations for CLIP–seq. Transcriptomic regions and gene classes were defined using annotations found in gencode v17. Depending on the analysis, clusters were associated by the Gencode-annotated 5′UTR, 3′UTR, CDS or intronic regions. If a cluster overlapped multiple regions, or a single part of a transcript was annotated as multiple regions, clusters were iteratively assigned first as CDS, then 3′UTR, 5′UTR and finally as proximal (<500 bases from an exon) or distal (>500 bases from an exon) introns. Overlapping peaks were calculated using bedtools and pybedtools[35,36].

Gene ontology analysis for CLIP–seq. Significantly enriched gene ontology (GO) terms were identified using a hypergeometric test that compared the number of genes that were MSI2 targets in each GO term to genes expressed in each GO term as the proper background. Expressed genes were identified using the control samples in SRA study SRP012062. Mapping was performed identically to CLIP–seq mapping, without peak calling and changing the STAR parameter outFilterMultimapNmax to 10. Counts were calculated with featureCounts[37] and RPKMs were then computed. Only genes with a mean RPKM > 1 between the two samples were used in the background expressed set.

De novo motif and conservation analysis for CLIP–seq. Randomly located clusters within the same genic regions as predicted MSI2 clusters were used to calculate a background distribution for motif and conservation analyses. Motif analysis was performed using the HOMER algorithm as in ref. 34. For evolutionary sequence conservation analysis, the mean (mammalian) phastCons score for each cluster was used.

Immunofluorescence. CD34+ cells (>5 × 10^4) were transduced with an MSI2-overexpression or control lentivirus. Three days later, GFP+ cells were sorted and then put back in to StemSpan medium containing growth factors IL-6 (20 ng ml^−1), SCF (100 ng ml^−1), FLT3-L (100 ng ml^−1) and TPO (20 ng ml^−1). A minimum of 10,000 cells were used for immunostaining at culture days 3 and 7 after GFP sorting. Cells were fixed in 2% PFA for 10 min, washed with PBS and then cytospun on to glass slides. Cytospun cells were then permeabilized (PBS, 0.2% Triton X-100) for 20 min, blocked (PBS, 0.1% saponin, 10% donkey serum) for 30 min and stained with primary antibodies (CYP1B1 (EPR14972, Abcam); HSP90 (68/hsp90, BD Biosciences)) in PBS with 10% donkey serum for 1 h. Detection with secondary antibody was performed in PBS 10% donkey serum with Alexa-647 donkey anti-rabbit antibody or Alexa-647 donkey anti-mouse antibodies for 45 min. Slides were mounted with Prolong Gold Antifade containing DAPI (Invitrogen). Several images (200–1,000 cells total) were captured per slide at 20× magnification using an Operetta HCS Reader (Perkin Elmer) with epifluorescence illumination and standard filter sets. Columbus software (Perkin Elmer) was used to automate the identification of nuclei and cytoplasm boundaries in order to quantify mean cell fluorescence.

Luciferase reporter gene assay. A 271-bp region of the *CYP1B1* 3′UTR that flanked CLIP–seq-identified MSI2-binding sites was cloned from human HEK293FT genomic DNA using the forward primer GTGACACAACTGTGTGATTAAAAGG and reverse primer TGATTTTTATTATTTTGGT AATGGTG and placed downstream of renilla luciferase in the dual-luciferase reporter vector pGL4 (Promega). A 271-bp geneblock (IDT) with 6 TAG > TCC mutations was cloned in to pGL4 using XbaI and NotI. The *HSP90* 3′UTR was amplified from HEK293FT genomic DNA with the forward primer TCTCTGGCTGAGGGATGACT and reverse primer TTTTAAGGCCAAGGAATTAAGTGA and cloned into pGL4. A geneblock of the *HSP90* 3′UTR (IDT) with 14 TAG > TCC mutations was cloned in to pGL4 using SfaAI and NotI. Co-transfection of wild-type or mutant luciferase reporter (40 ng) and control or MSI2-overexpressing lentiviral expression vector (100 ng) was performed in the NIH-3T3 cell line, which does not express MSI1 or
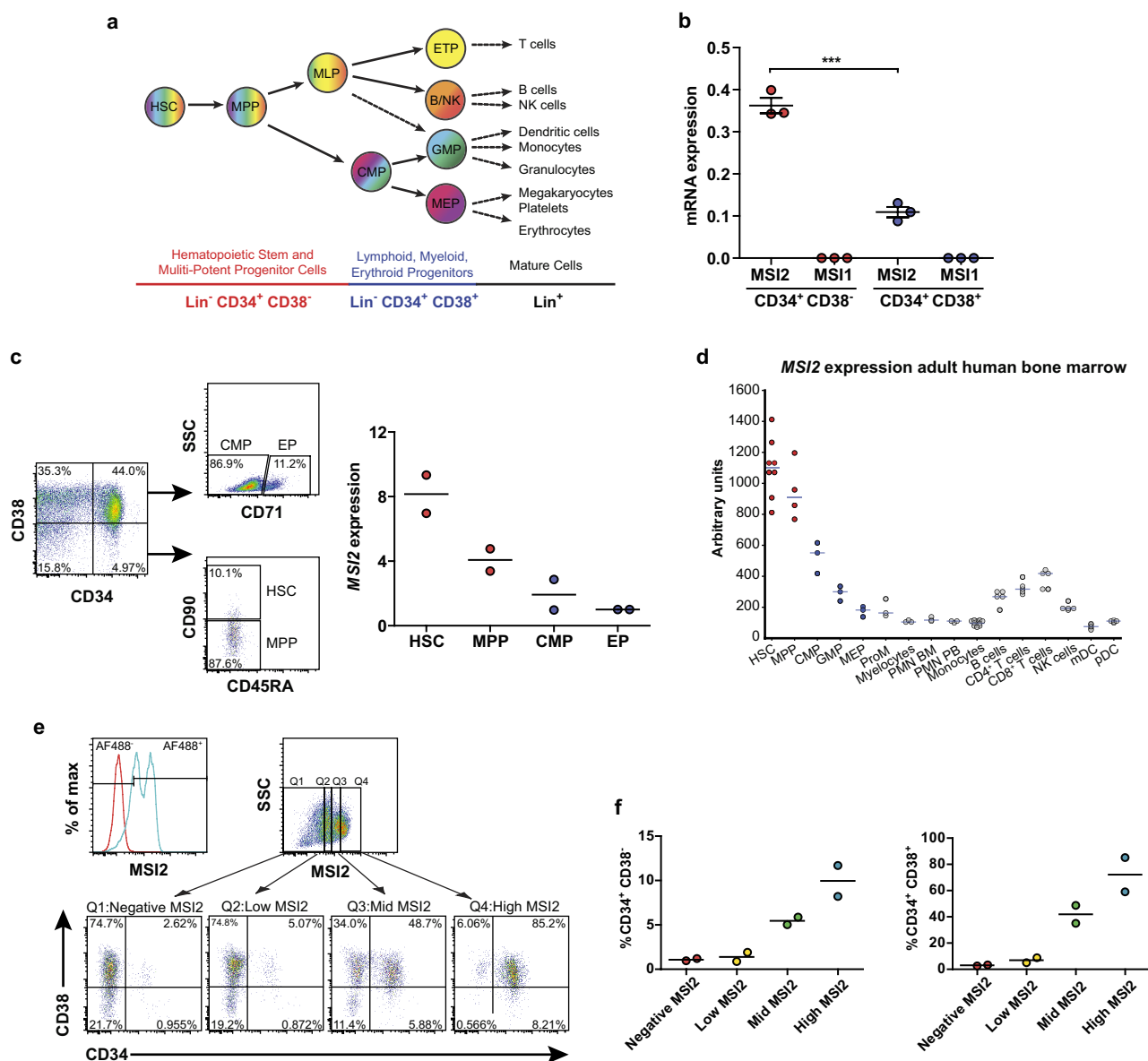
MSI2 (50,000 cells per co-transfection). Reporter activity was measured using the Dual-Luciferase Reporter Assay System (Promega) 36–40 h later.

MSI2-overexpressing suspension cultures with the AHR antagonist SR1 and agonist FICZ. For MSI2-overexpressing cultures with the AHR antagonist SR1, Lin− CD34+ cells were transduced with MSI2-overexpression or control lentivirus in medium supplemented with SR1 (750 nM; Abcam) or DMSO vehicle (0.1%). GFP+ cells were isolated (20,000 cells per culture) and allowed to proliferate with or without SR1 for an additional 7 days at which point they were counted and immunophenotyped for CD34 and CD133 expression. For MSI2-overexpressing cultures with the AHR agonist FICZ, Lin− CD34+ cells were transduced with MSI2-overexpression or control lentivirus. GFP+ cells were isolated (20,000 cells per culture) and allowed to proliferate with FICZ (200 nM; Santa Cruz Biotechnology) or DMSO (0.1%) for an additional 3 days, at which point they were immunophenotyped for CD34 and CD133 expression.

HSPC expansion with (E)-2,3′,4,5′-tetramethoxystilbene (TMS). Lin− CD34+ cells were cultured for 72 h (lentiviral treated but non-transduced flow-sorted GFP− cells) in StemSpan medium containing growth factors IL-6 (20 ng ml^−1), SCF (100 ng ml^−1), FLT3-L (100 ng ml^−1) and TPO (20 ng ml^−1) before the addition of the CYP1B1 inhibitor TMS (Abcam) at a concentration of 10 μM or mock treatment with 0.1% DMSO. Equal numbers of cells (12,000 per condition) were then allowed to proliferate for 7 days at which point they were counted and immunophenotyped for CD34 and CD133 expression.
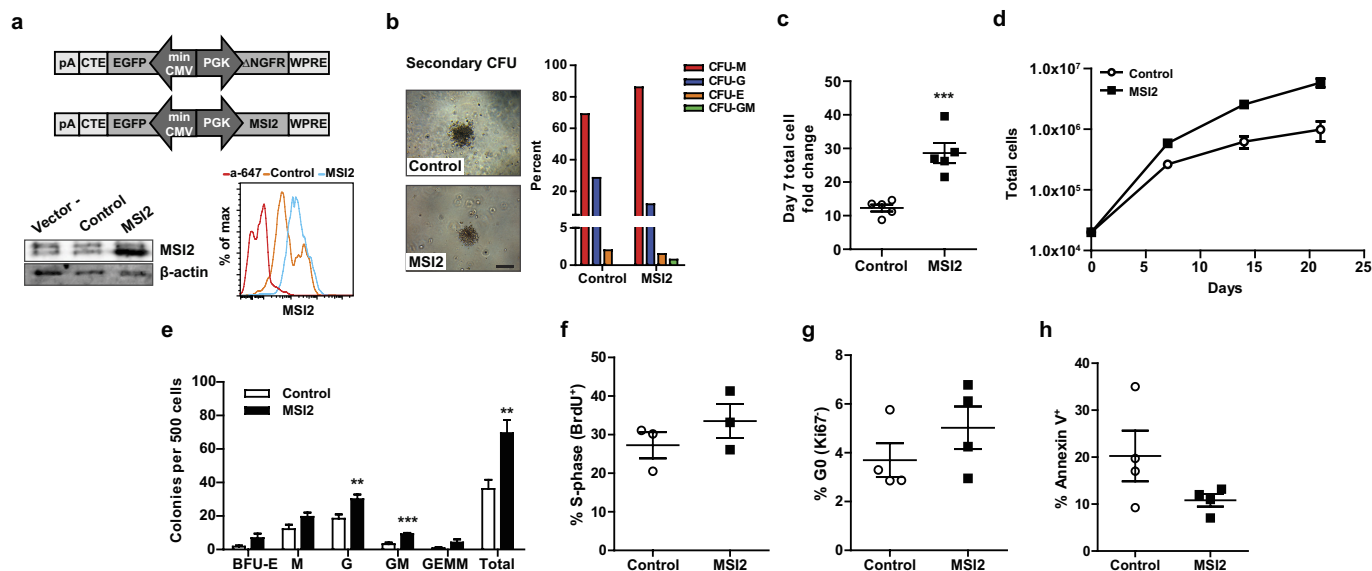
Statistical analysis. Unless stated otherwise (that is, analysis of RNA–seq and CLIP–seq data sets), all statistical analysis was performed using GraphPad Prism (GraphPad Software version 5.0). Unpaired student *t*-tests or Mann–Whitney tests were performed with $P < 0.05$ as the cut-off for statistical significance. No statistical methods were used to predetermine sample size.

18. Doulatov, S. *et al.* PLZF is a regulator of homeostatic and cytokine-induced myeloid development. *Genes Dev.* **23,** 2076–2087 (2009).
19. Majeti, R., Park, C. Y. & Weissman, I. L. Identification of a hierarchy of multipotent hematopoietic progenitors in human cord blood. *Cell Stem Cell* **1,** 635–645 (2007).
20. Bagger, F. O. *et al.* HemaExplorer: a database of mRNA expression profiles in normal and malignant haematopoiesis. *Nucleic Acids Res.* **41,** D1034–D1039 (2013).
21. Amendola, M., Venneri, M. A., Biffi, A., Vigna, E. & Naldini, L. Coordinate dual-gene transgenesis by lentiviral vectors carrying synthetic bidirectional promoters. *Nature Biotechnol.* **23,** 108–116 (2005).
22. van Galen, P. *et al.* The unfolded protein response governs integrity of the haematopoietic stem-cell pool during stress. *Nature* **510,** 268–272 (2014).
23. Lechman, E. R. *et al.* Attenuation of miR-126 activity expands HSC in vivo without exhaustion. *Cell Stem Cell* **11,** 799–811 (2012).
24. Carow, C. E., Hangoc, G. & Broxmeyer, H. E. Human multipotential progenitor cells (CFU-GEMM) have extensive replating capacity for secondary CFU-GEMM: an effect enhanced by cord blood plasma. *Blood* **81,** 942–949 (1993).
25. Milyavsky, M. *et al.* A distinctive DNA damage response in human hematopoietic stem cells reveals an apoptosis-independent role for p53 in self-renewal. *Cell Stem Cell* **7,** 186–197 (2010).
26. Janky, R. *et al.* iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLOS Comput. Biol.* **10,** e1003731 (2014).
27. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
28. Kwon, A. T., Arenillas, D. J., Worsley Hunt, R. & Wasserman, W. W. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3* **2,** 987–1002 (2012).
29. Colvin, G. A. *et al.* Murine marrow cellularity and the concept of stem cell competition: geographic and quantitative determinants in stem cell biology. *Leukemia* **18,** 575–583 (2004).
30. Hu, Y. & Smyth, G. K. ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J. Immunol. Methods* **347,** 70–78 (2009).
31. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
32. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
33. Darnell, R. CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. *Cold Spring Harb. Protoc.* **2012,** 1146–1160 (2012).
34. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Struct. Mol. Biol.* **20,** 1434–1442 (2013).
35. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
36. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27,** 3423–3424 (2011).
37. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).
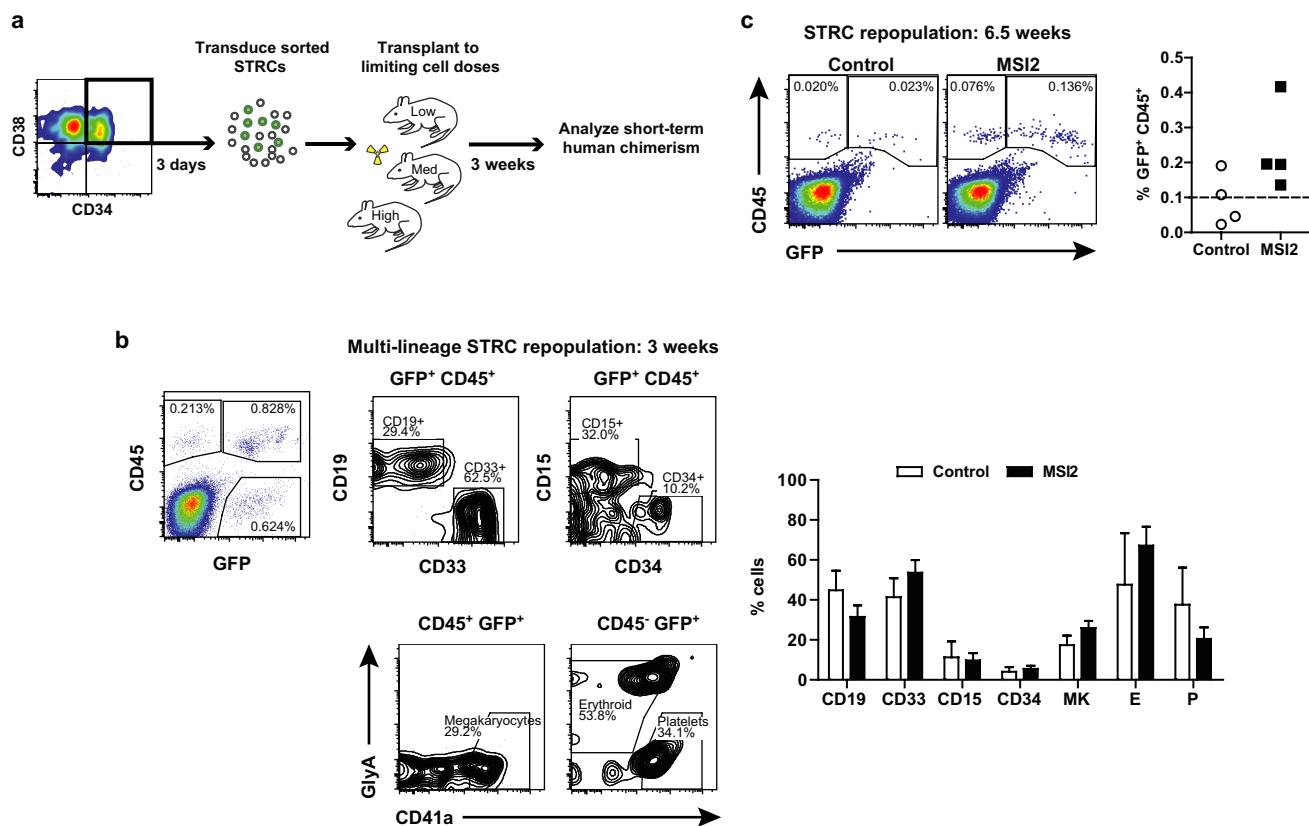
Extended Data Figure 1 | MSI2 is highly expressed in human haematopoietic stem and progenitor cell populations. a, Schematic of the human haematopoietic hierarchy showing key primitive cell populations and simplified surface marker expression. **b**, qRT–PCR analysis of *MSI1* and *MSI2* expression in Lin$^-$ cord blood (CB) cell populations ($n = 3$ independent Lin$^-$ CB samples). **c**, Gating strategy used to sort sub-fractions of Lin$^-$ CB HSPCs for *MSI2* qRT–PCR expression analysis ($n = 2$ independent pooled Lin$^-$ CB samples). **d**, *MSI2* expression across the human haematopoietic hierarchy. Each circle represents an

independent gene expression data set curated by HemaExplorer. **e**, Intracellular flow cytometry analysis of MSI2 protein levels in Lin$^-$ CB. Histograms show background staining with secondary antibody (red) and positive staining with anti-MSI2 antibody plus secondary in Lin$^-$ CB (blue). MSI2 fluorescence intensity was divided into quartiles of negative (Q1), low (Q2), mid (Q3) and high (Q4) level expression. **f**, Plots show cell percentage within each quartile from **e** that are CD34$^+$ CD38$^-$ (left) and CD34$^+$ CD38$^+$ (right) ($n = 2$ independent Lin$^-$ CB samples). All data presented as mean ± s.e.m. Unpaired *t*-test, *$P < 0.05$; ***$P < 0.001$.
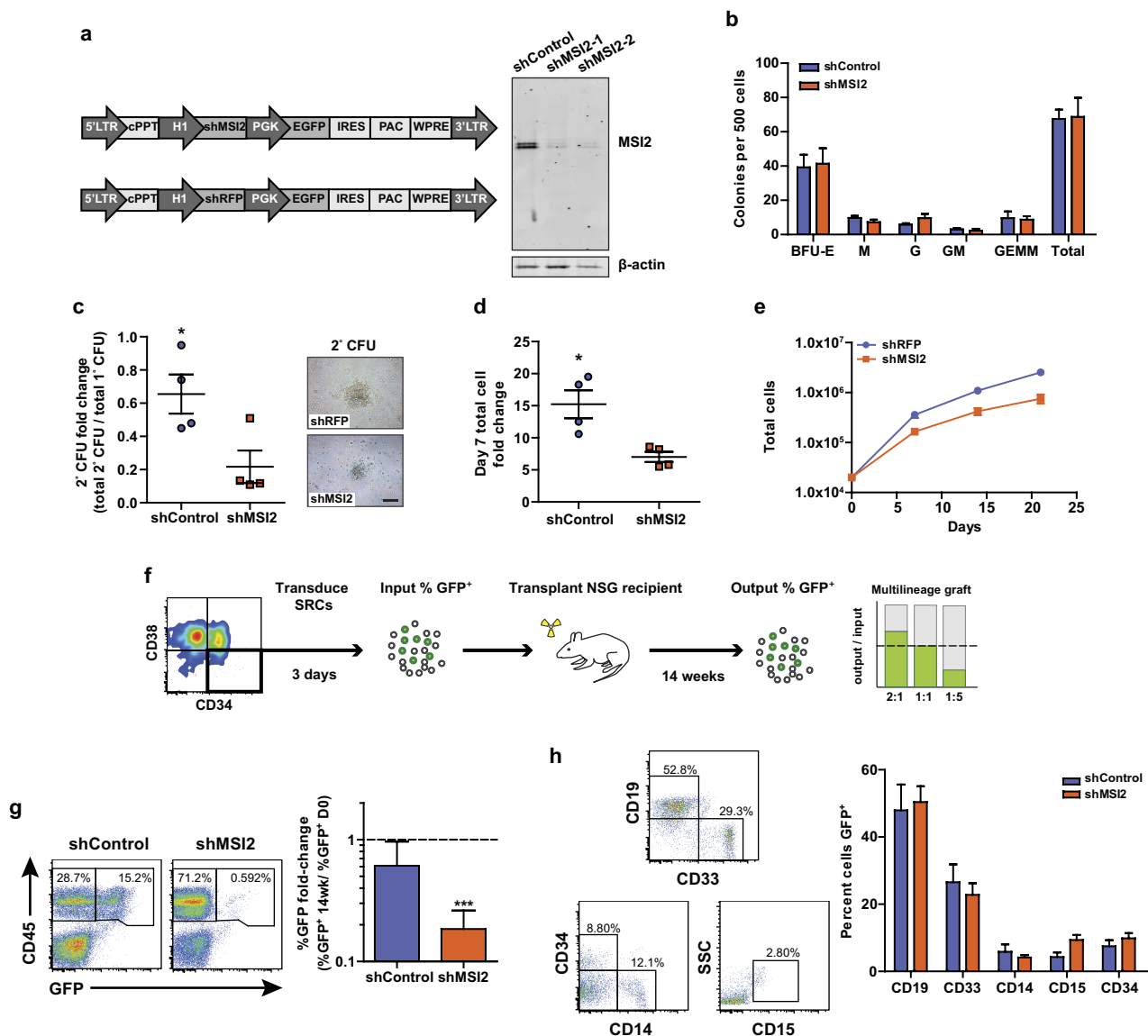
Extended Data Figure 2 | MSI2 overexpression enhances *in vitro* culture of primitive CB cells. a, Top: schematic of bi-directional promoter lentivirus used to overexpress MSI2. Bottom: western blot and histogram showing intracellular flow validation of enforced MSI2 expression in 293FT cells (left) and Lin⁻ CB (right), respectively. **b**, Representative images of secondary CFU made from replated control and MSI2-overexpressing (MSI2) CFU-GEMMs and types of colonies made. Scale bar, 200 μm. **c**, Fold change in Lin⁻ CB transduced cell number after 7 days in culture following transduction ($n = 5$ experiments). **d**, Growth curve over 21 days of transduced Lin⁻ CB cells ($n = 4$ experiments). **e**, Colony output of transduced Lin⁻ CB from day 7 cultures ($n = 8$ cultures from 4 experiments). **f**, BrdU cell cycle analysis of transduced Lin⁻ CB cells from day 10 cultures ($n = 3$ experiments). **g**, Ki67 cell cycle analysis of transduced Lin⁻ CB cells from day 4 cultures ($n = 4$ experiments). **h**, Apoptotic and dead cells in day 7 cultures of transduced Lin⁻ CB by Annexin V staining ($n = 3$ experiments). Western blot source data are available in Supplementary Fig. 1. All data presented as mean ± s.e.m. Unpaired *t*-test, **$P < 0.01$; ***$P < 0.001$.
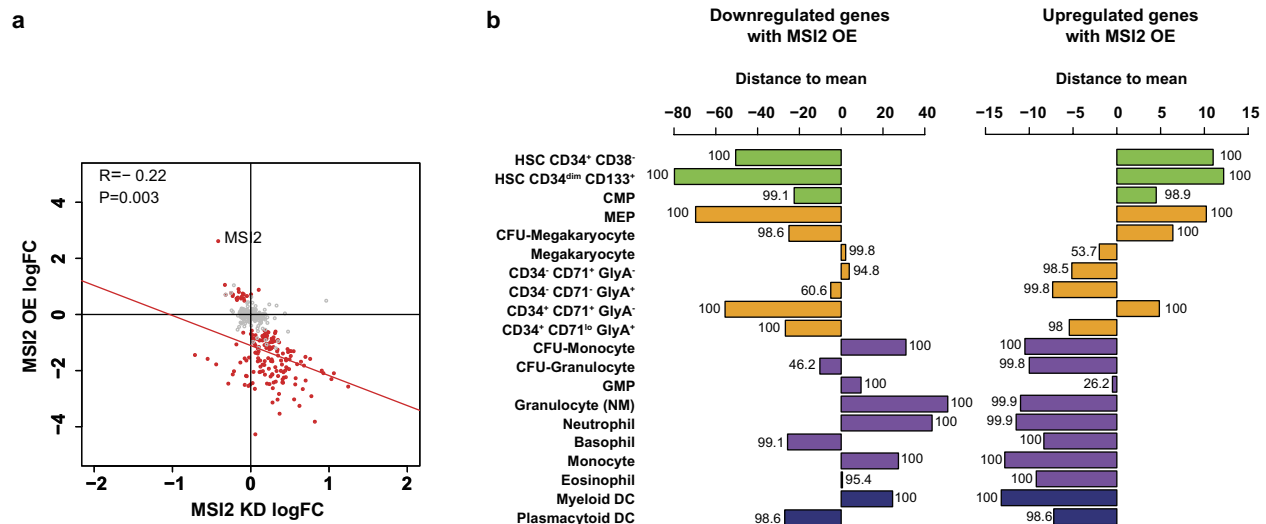
Extended Data Figure 3 | MSI2 overexpression does not affect STRC lineage output and extends STRC-mediated engraftment. a, Schematic of STRC LDA experimental setup. **b**, Left: gating strategy to identify engrafted GFP$^+$ CD45$^+$ progenitor and myelo-lympho lineage-positive cell types or GFP$^+$ CD45$^-$ erythroid cells and platelets. Right: summary of lineage output in the injected femur 3 weeks after transplantation

($n = 4$ mice for control and $n = 18$ mice for MSI2 overexpressing cells). MK, megakaryocyte; E, erythroid cells; P, platelets. **c**, Representative flow plots and summary of transduced STRC read out for engraftment with human CD45$^+$ cells at 6.5 weeks post-transplant ($n = 4$ mice per condition). All data presented as mean $\pm$ s.e.m.
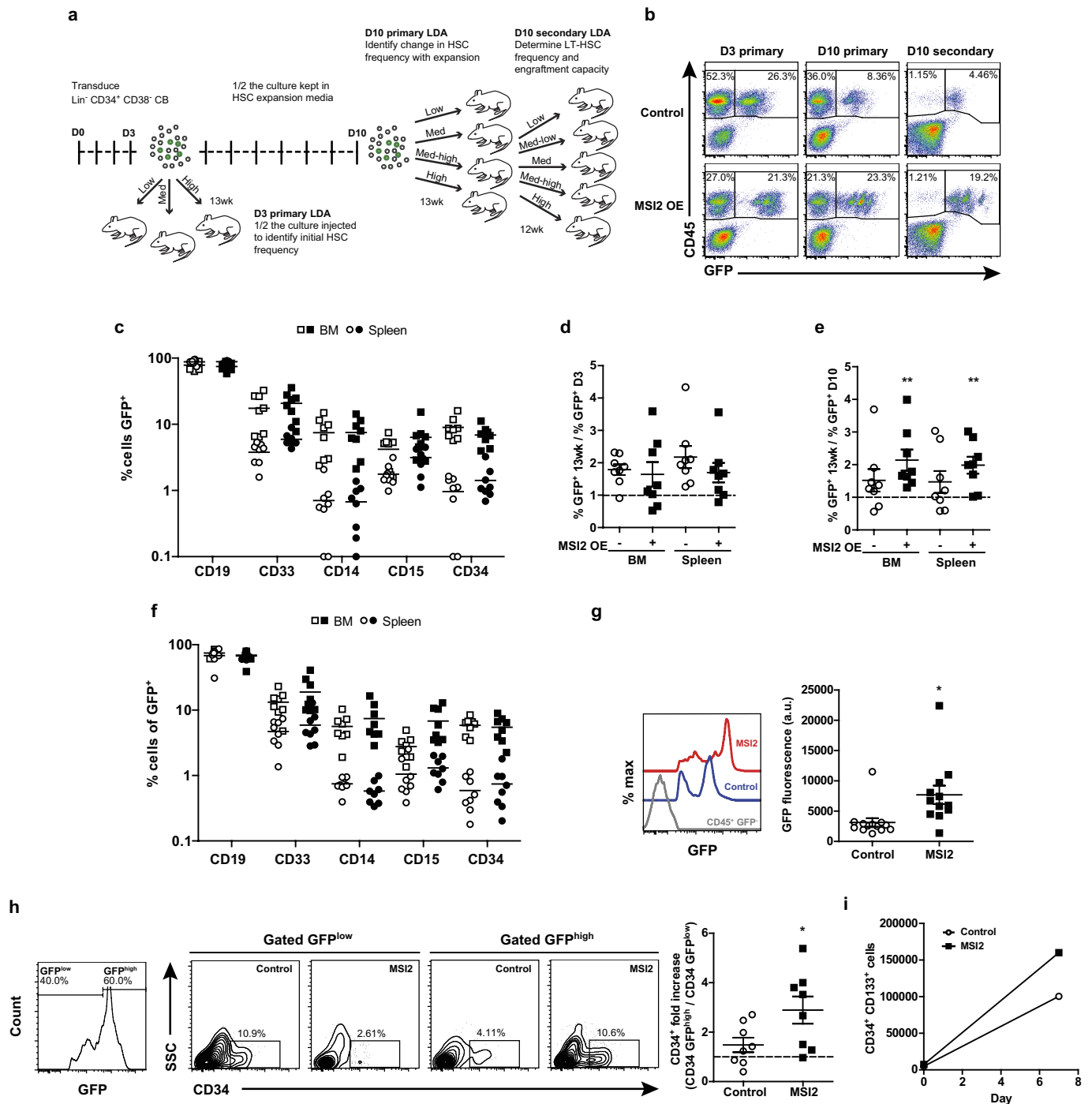
Extended Data Figure 4 | MSI2 knockdown impairs secondary CFU replating potential and HSC engraftment capacity. a, Left: schematic of MSI2- and control RFP-targeted shRNA lentiviruses. Right: confirmation of MSI2 protein knockdown (both isoforms that can be detected by western blot) in transduced NB4 cells. **b**, CFU production by shMSI2- and shControl-transduced Lin$^-$ CB ($n = 8$ cultures from 4 experiments). **c**, Secondary CFU output from shMSI2-transduced Lin$^-$ CB and images of representative secondary CFUs (scale bar, 200 μm; performed on $n = 4$ cultures from 2 experiments). **d**, Fold change in transduced cell number after 7 days in culture ($n = 4$ experiments). **e**, Growth curves of cultures initiated with transduced Lin$^-$ CB cells ($n = 4$ experiments). **f**, Experimental design to read out changes in HSC capacity with MSI2

knockdown. **g**, Left: representative flow analysis of transduced CD34$^+$ CD38$^-$-derived human chimaerism in NSG mouse bone marrow. Right: ratio of the percentage of GFP$^+$ cells in the CD45$^+$ population post-transplant to the initial pre-transplant GFP$^+$ cell percentage. Dotted line indicates that the proportion of GFP$^+$ cells is unchanged relative to input. One sample t-test, no change $= 1$; $n = 6$ mice receiving shControl and $n = 8$ mice receiving shMSI2-transduced cells pooled from two experiments. **h**, Representative flow plots and summary of multilineage engraftment with shControl and shMSI2 cells (gated on GFP$^+$ cells). Western blot source data are shown in Supplementary Fig. 1. Data presented as mean ± s.e.m. Unpaired t-test, *$P < 0.05$; ***$P < 0.001$.
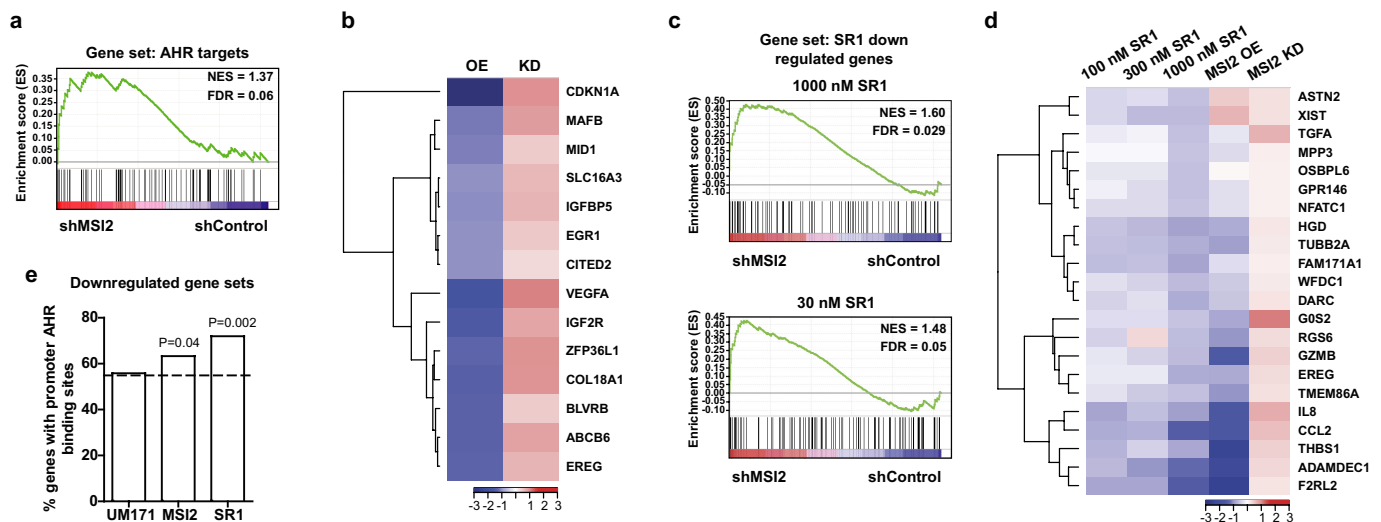
Extended Data Figure 5 | MSI2 overexpression confers an HSC gene expression signature. a, Genes that are upregulated (21 genes, logFC >0) or downregulated (156 genes, logFC <0) in MSI2-overexpressing (OE) cells relative to control cells with FDR < 0.05 were compared to expression data from MSI2 knockdown cells normalized to shControl expression data. Red circles represent 177 genes that were significantly differentially expressed in MSI2-overexpressing cells. Gray outlined circles represent random genes (equal number of grey circles and red circles). Only genes that were significantly up- or downregulated in MSI2-overexpressing cells showed anti-correlation with MSI2 knockdown cells. **b**, Genes that were differentially expressed between MSI2-overexpressing and control cells (FDR < 0.05) compared to DMAP populations. Numbers beside each bar indicate the percentage of time for which the observed value (set of up- or downregulated genes) was better represented in that population than random values (equal number of randomly selected genes based on 1,000 trials).
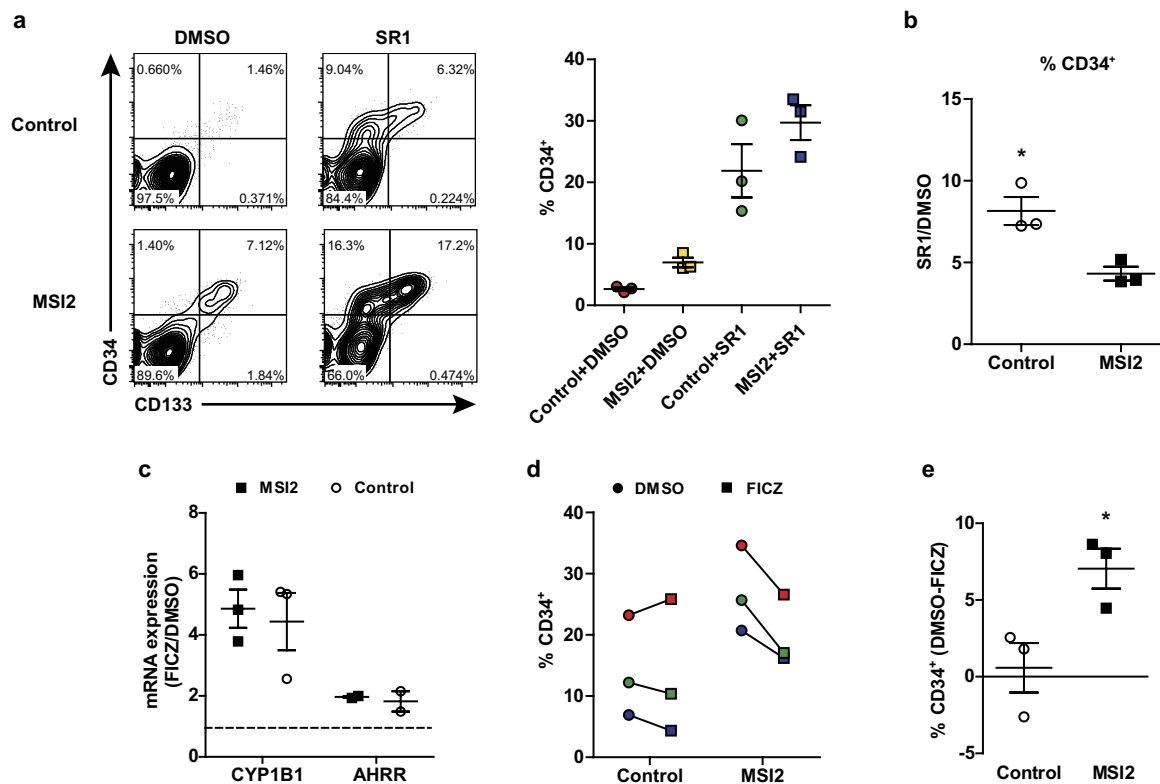
Extended Data Figure 6 | MSI2 overexpression enhances HSC activity after *ex vivo* culture. a, Experimental procedure for measuring changes in HSC engraftment capacity and frequency with *ex vivo* culture. **b**, Representative flow plots of CD45[+] GFP[+] reconstitution from mice receiving the highest cell dose transplanted per time point. **c**, Multilineage engraftment of mice injected with D3 cultures. **d**, Proportion of the human CD45[+] graft containing GFP[+] cells from mice receiving the two highest doses of D3 primary grafts relative to pre-transplant levels of GFP[+] cells before expansion (*n* = 8 mice for each dose). **e**, Proportion of the human CD45[+] graft containing GFP[+] cells from mice receiving the two highest doses of D10 primary grafts relative to pre-transplant levels of GFP[+] cells after expansion (*n* = 8 mice for each dose, one-sample *t*-test, no change = 1). **f**, Multilineage engraftment of mice injected with D10 cultures. **g**, GFP mean fluoresence intensity (MFI) in D10 primary cell-engrafted mice. Data are from mice transplanted with the highest three doses; *n* = 11 control and 13 MSI2-overexpressing cell-engrafted mice. **h**, CD34 expression in GFP[high] (top 60%) relative to GFP[low] (bottom 40%) gated cells (set per mouse) from engrafted recipients in **e**. **i**, Number of transduced phenotyped HSCs after 7 days of culture from HSC expansion experiment described in **a**. Symbols represent individual mice and shaded symbols represent mice grafted with MSI2-overexpressing cells. All data presented as mean ± s.e.m. Unpaired *t*-test, *P < 0.05.
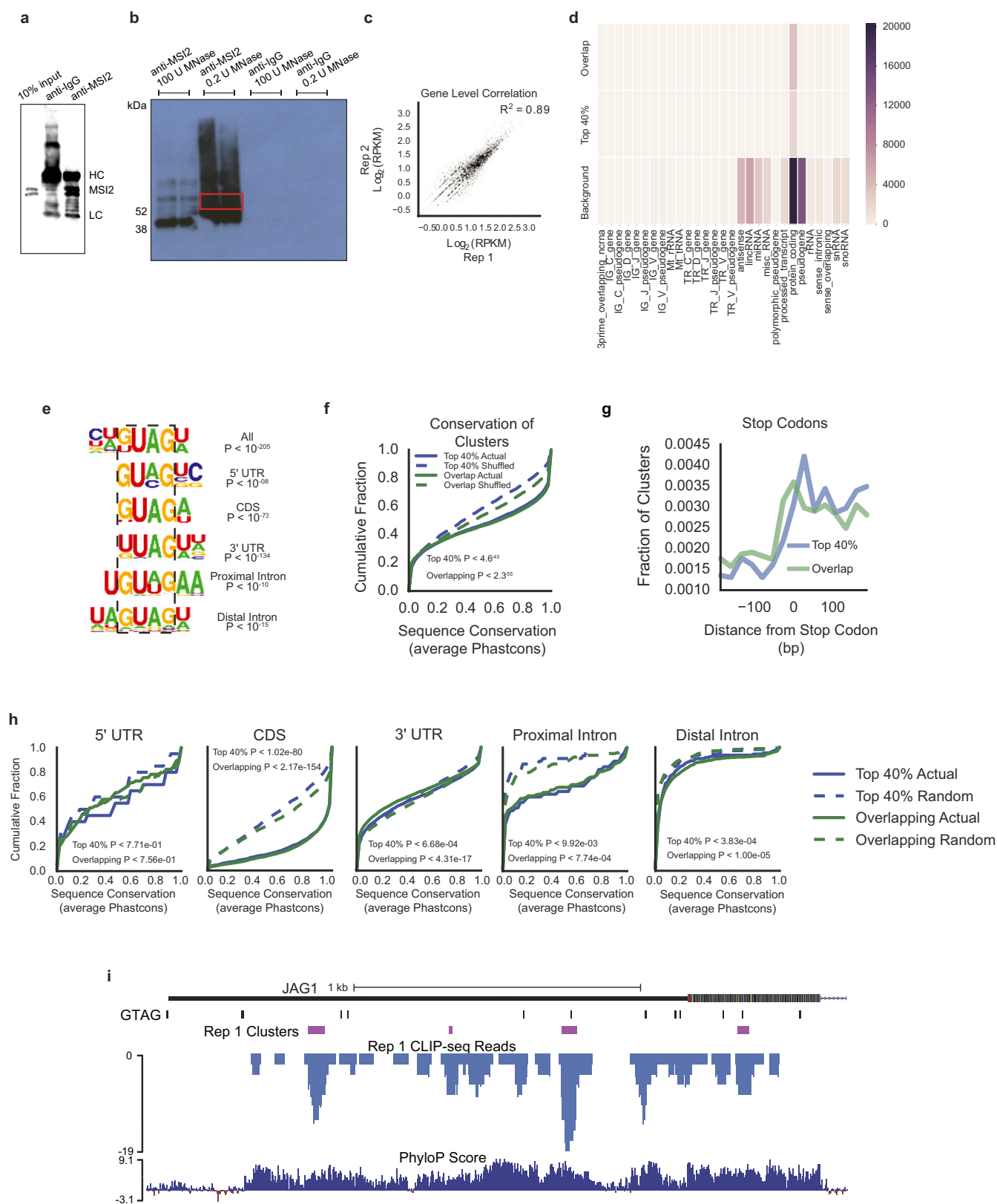
Extended Data Figure 7 | Predicted AHR targets and genes downregulated by SR1 or MSI2 overexpression are upregulated by MSI2 knockdown. a, Predicted AHR targets were identified with the iRegulon tool and compared with MSI2 knockdown normalized to shControl-upregulated gene signature by GSEA. **b**, log fold-change of MSI2-overexpression and knockdown shared leading edge AHR target genes from GSEA. **c**, GSEA comparing gene sets downregulated by SR1 high and low dose with the MSI2 knockdown upregulated gene signature.

d, Heatmap and log fold-change of shared leading edge genes identified by GSEA from MSI2 overexpression, MSI2 knockdown and SR1 at varying doses. **e**, The percentage of downregulated genes in UM171-treated, SR1-treated and MSI2-overexpressing cells containing at least one AHR-binding site within 1,500 bp upstream or downstream of the transcription start site. Dotted line indicates the background percentage of genes with AHR-binding sites. *P* values were generated relative to background with Fisher's exact test.

Extended Data Figure 8 | AHR antagonism with SR1 has redundant effects with MSI2 overexpression, and AHR activation with MSI2 overexpression results in a loss of HSPCs. a, Representative flow plots and summary of CD34 expression in MSI2-overexpressing and control transduced CD34$^+$ CB cells grown for 10 days in medium containing SR1 or DMSO vehicle ($n = 3$ experiments). **b**, Fold change in CD34 expression from results shown in **a. c**, Fold increase in *CYP1B1* and *AHRR* transcript lev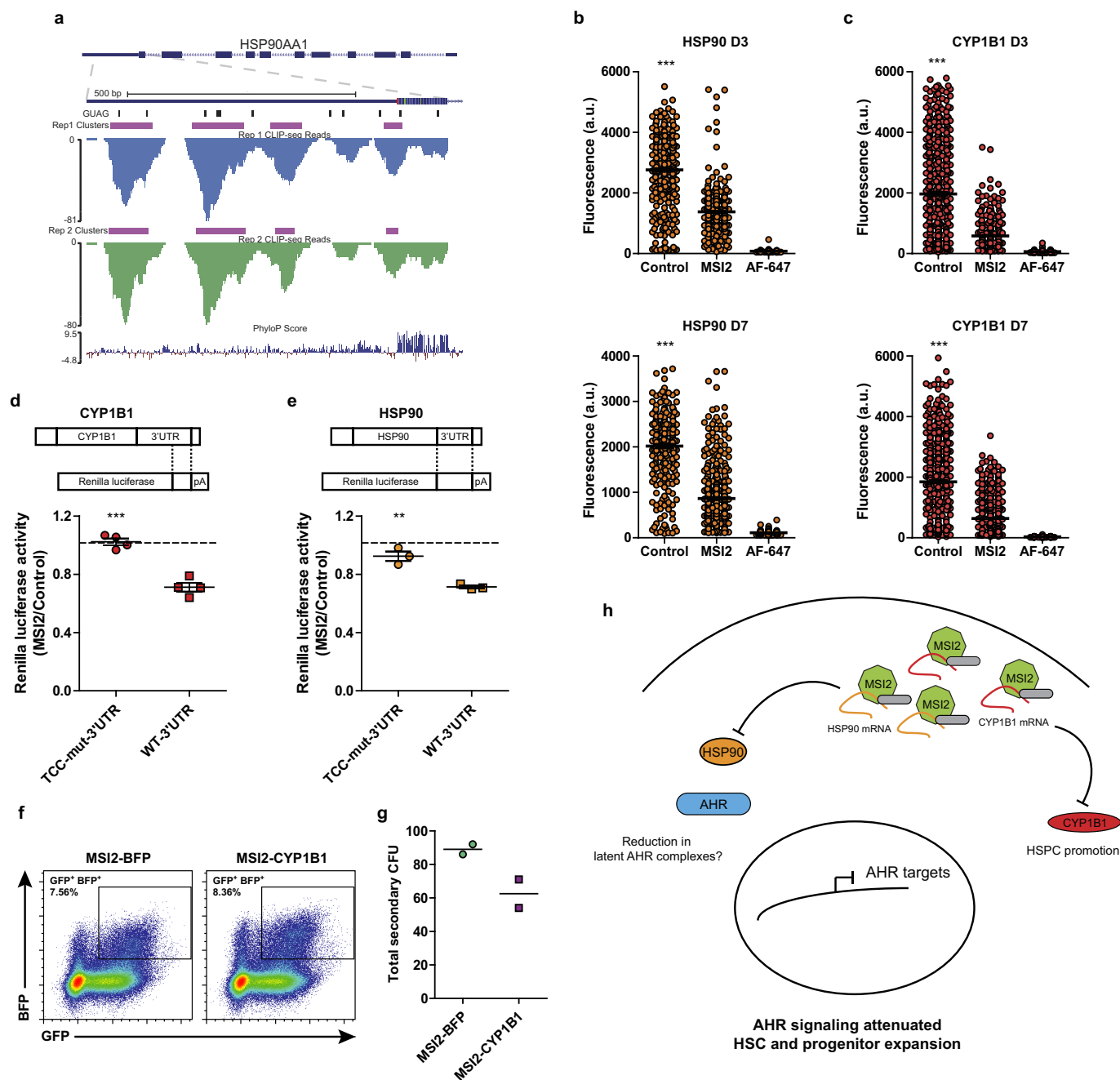els after FICZ treatment in transduced cultures ($n = 3$ experiments). **d**, Transduced CD34$^+$ CB cells grown for 3 days in medium supplemented with FICZ and the corresponding change in CD34 expression. Each coloured pair (DMSO and FICZ) represents a matched CB sample ($n = 3$ experiments). **e**, Differences in culture CD34 levels from **d**. All data presented as mean ± s.e.m. Unpaired *t*-test, *$P < 0.05$.

Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | MSI2 preferentially binds mature mRNA within the 3′UTR. a, Validation of the capacity of the anti-MSI2 antibody to immunoprecipitate MSI2 compared to IgG control pulldowns (heavy chain, HC; light chain, LC). **b,** Autoradiogram showing anti-MSI2 immunoprecipitated, MNase digested and radiolabelled RNA isolated for CLIP library construction and sequencing (red box). Low levels of MNase show a smearing pattern extending upwards from the modal weight of MSI2. **c,** Scatter plot of total number of uniquely mapped CLIP–seq reads for each gene, comparing both replicates. **d,** Heatmap indicating the number of different classes of Gencode-annotated genes that contain at least one predicted MSI-binding site. **e,** Consensus motifs within MSI2 clusters in the different genic regions. *P* values for the most statistically significant enriched motif are presented for all overlapping clusters between replicates. **f,** Cumulative distribution function of mean conservation score (Phastcons) of MSI2 clusters, compared to a shuffled background control, computed for all overlapping clusters and the top 40% of overlapping clusters. *P* values were obtained

by a Kolgomorov–Smirnov two-tailed test comparing the distributions from actual and shuffled locations. **g,** Number of clusters within 200 bases of the annotated stop codon in known mRNA transcripts for all overlapping clusters between replicates and the top 40% of overlapping clusters. **h,** Cumulative distribution function of mean conservation score (Phastcons) of MSI2 clusters, compared to a shuffled background control, computed for overlapping clusters between the replicates and the top 40% of overlapping clusters found in different genic regions. Similarity between the 3′UTR conservation for the top 40% and the shuffled background control is probably due to MSI2 sites being small and not needing structural contexts for conservation. *P* values were obtained by a Kolgomorov–Smirnov two-tailed test comparing the distributions from actual and shuffled locations. **i,** Genome browser views displaying CLIP–seq mapped reads from replicate 1 (blue), predicted clusters (purple), exact matches for the GUAG sequence (black) and mammal conservation scores (PhyloP) in the 3′UTRs for a previously predicted Msi1 target.

Extended Data Figure 10 | MSI2 overexpression represses CYP1B1 and HSP90 3′UTR Renilla Luciferase reporter activity. a, CLIP–seq reads (replicate 1 in blue and replicate 2 in green) and clusters (purple) mapped to the 3′UTR of HSP90. Matches to the GUAG motif are shown in black. Mammal PhyloP score listed in last track. **b, c**, Representative data of mean per cell fluoresence for HSP90 and CYP1B1 protein in transduced CD34+ CB. Protein level in cells during *in vitro* culture was analysed 3 days (D3) and 7 days (D7) after transduction and sorting for GFP. Corresponding secondary-alone antibody staining is shown for each experiment. Each circle represents a cell, and more than 200 cells were analysed per condition. **d, e**, Levels of renilla luciferase activity in NIH-3T3 cells co-transfected with control or MSI2 overexpression vectors and the CYP1B1 or HSP90 wild-type or TCC mutant 3′UTR luciferase reporter (dotted

line indicates no change in renilla activity; $n = 4$ CYP1B1 and $n = 3$ HSP90 experiments). **f**, Flow plots of co-transduced CD34+ CB cells with MSI2 (GFP) and CYP1B1 (BFP) lentivirus. **g**, GFP+ BFP+ CFU-GEMMs generated from **f** were replated into secondary CFU assays and the total number of colonies formed was counted. A total of 24 CFU-GEMMs from MSI2-BFP and MSI2-CYP1B1 were replated ($n = 2$ experiments). Data presented as mean ± s.e.m. Unpaired *t*-test, ***$P < 0.001$, **$P < 0.01$. **h**, A model for AHR pathway attenuation by MSI2 post-transcriptional processing. MSI2 mediates the post-transcriptional downregulation of HSP90 at the outset of culture and continuously represses the prominent AHR pathway effector CYP1B1 to facilitate HSPC expansion. The resulting MSI2-mediated repression of AHR signalling enforces a self-renewal program and allows HSPC expansion *ex vivo*.