Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis

Victoria Moignard1, Iain C. Macaulay2, Gemma Swiers3, Florian Buettner4, Judith Schütte5, Fernando J. Calero-Nieto6, Sarah Kinston1, Anagha Joshi1, Rebecca Hannah1, Fabian J. Theis4,5, Sten Eirik Jacobsen2, Marella F. de Bruijn3 and Berthold Göttgens1,6

Haematopoiesis has long served as a model system for studying cell fate decisions during stem cell differentiation1,2. At the molecular level, transcription factors are major drivers of cellular identity and cell fate transitions as exemplified by their key role in reprogramming3 and lineage switching experiments4–6. Transcription factors function within wider regulatory networks, the connectivity of which can be revealed using classical transcriptional assays or inferred from global expression and transcription factor binding profiling studies7–12. However, these experiments report population averages, whereas cell fate choices are made by individual cells. The importance of studying single cells is emphasized by the known functional heterogeneity in haematopoietic stem cells (HSCs) as well as other cell types, which manifests as relatively stable subpopulations with either balanced production of myeloid and lymphoid cells or a deficiency in lymphoid potential13–15. HSCs have also been reported to be heterogeneous in gene expression16–18, although previous studies have been limited in terms of numbers of genes, cells or populations analysed. Recent advances in microfluidics technologies have facilitated high-throughput quantitative real-time PCR analysis of tens of genes in hundreds of single cells simultaneously19. This technology has recently been used to resolve cell populations in 64 cell mouse embryos20, to dissect cellular heterogeneity in human colon cancer21 and to reveal expression variation in myeloid precursor cells, which resolved following erythroid commitment22.

Here we analysed expression of a network of 18 densely interconnected transcription factors in 597 single cells from five primary haematopoietic stem and progenitor cell populations, which not only revealed characteristic expression states for the different cell populations, but also identified previously unrecognized regulatory relationships. This included a putative regulatory triad consisting of Gata2, Gfi1 and Gfi1b, which was validated using cell line and transgenic mouse assays. Our findings suggest that GATA2 may function in a regulatory loop to modulate Gfi1/Gfi1b cross-antagonism during entry into the myelolymphoid lineages, thus demonstrating that high-throughput

1 Department of Haematology, Wellcome Trust and MRC Cambridge Stem Cell Institute and Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK. 2Haematopoietic Stem Cell Laboratory, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK. 3MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK. 4Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. 5Department of Mathematics, Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany. 6Correspondence should be addressed to B.G. (e-mail: bg200@cam.ac.uk)

Received 24 June 2012; accepted 8 February 2013; published online 24 March 2013; corrected online 2 April 2013; DOI: 10.1038/ncb2709
Figure 1 Single-cell gene expression analysis of a core haematopoietic transcriptional regulatory network. (a) Schematic of the haematopoietic hierarchy, with the megakaryocyte-erythroid lineage in red, the myeloid lineages in orange, and the lymphoid lineage in blue. Cell types investigated in this study are outlined in the colours used to represent these populations in subsequent figures, and encompass both early multipotent stem and progenitors and committed progenitors for each of the main haematopoietic lineages. Cell surface phenotypes were LSK CD150⁺CD48⁻HSC (also gated as CD34⁺Flt3⁻), LSK Flt3⁺LMPP, Lin⁻IL7Ra⁻Kit⁺Sca-1⁻CMP, CD41⁺CD16/32hi GMP (also gated Lin⁻c-Kit⁺CD150⁻) and CD16/32⁻CD41⁻CD150⁻CD105⁺PreMegE (also gated Lin⁻c-Kit⁺). LT-HSC, long-term haematopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; GMP, granulocyte–monocyte progenitor; PreMegE, megakaryocyte-erythroid progenitor; NK cell, natural killer cell. (b) Network diagram of data curated from the literature and protein interaction databases (STRING and FunctionalNet) illustrating the complex interactions between 18 core haematopoietic transcription factors. Green lines indicate functional relationships and red lines indicate direct protein–protein interactions. Activating and inhibitory connections are not distinguished.

single-cell transcription factor expression analysis provides a powerful approach towards the identification of regulatory network links.

RESULTS

Single-cell expression analysis reveals heterogeneity in transcription factor expression in haematopoietic stem and progenitor cells

To study core regulatory circuits during early haematopoietic differentiation stages, we performed gene expression analysis for transcription factors in single primary haematopoietic stem/progenitor cells prospectively isolated from mouse bone marrow by fluorescence activated cell sorting (FACS). We analysed long-term haematopoietic stem cells (LSK CD150⁺CD48⁻HSC; ref. 23), lymphoid-primed multipotent progenitors (LSK Flt3⁺LMPP; ref. 24), bipotential megakaryocyte–erythroid progenitors (CD16/32⁻CD41⁺CD150⁻CD105⁺PreMegE; ref. 25), granulocyte–monocyte progenitors (CD41⁻CD16/32⁺GMPs; ref. 25,26) and common lymphoid progenitors (Lin⁻IL7Ra⁻Kit⁺Sca-1⁻CLPs; ref. 27, Fig. 1a and Supplementary Fig. S1). A total of 597 single cells (123 CLPs, 124 GMPs, 121 HSCs, 116 LMPPs and 113 PreMegEs) passed quality-control measures (see Methods).

Single-cell gene expression analysis was performed for 24 genes in all 597 cells (see Supplementary Table S3 for raw Ct data). Our gene set included 18 transcription factors (Fig. 1b) with known key roles in haematopoiesis, as well as five housekeeping genes and the stem cell factor receptor c-Kit, which is expressed on the surface of all analysed haematopoietic stem/progenitor subsets²³–²⁸. We have previously reported the potential for abundant regulatory linkages between many of the 18 transcription factors, or have examined a lineage-specific set of genes²² and a similarly densely interconnected network was obtained using data curated from the literature and protein interaction databases (Fig. 1b). Importantly, although previous studies have examined individual HSCs and commitment to the erythroid lineage, they have been limited in cell numbers, were focused on expression heterogeneity, or have examined a lineage-specific set of genes. Moreover, the potential for identifying regulatory connections from single-cell gene expression profiling had not been demonstrated, nor had the potential for dynamic changes of regulatory network states been studied during the differentiation of HSCs into the various multipotent blood progenitors.

Single-cell gene expression analysis recovered expected expression patterns for the 18 transcription factors as well as housekeepers and c-Kit (Fig. 2). For example, the c-Kit expression level was highest in HSCs and gradually reduced in the progenitor populations, consistent with the reported downregulation in progenitors. Gata1 is known to be expressed at high levels in erythroid and megakaryocyte...
Figure 2 Haematopoietic transcription factors show heterogeneous expression in haematopoietic stem and progenitor cells. Density plots for 18 transcription factors, the stem cell factor receptor c-Kit and the housekeeping gene Ubc in five haematopoietic stem and progenitor lineages, but not in HSCs (ref. 34), and here was expressed in around two-thirds of PreMegE cells, yet absent in almost all cells of the other populations. Likewise, Gata2 is known to be expressed in HSCs and during megakaryopoiesis\(^ {35,36} \), and in our data was expressed in most HSCs and PreMegEs but at lower levels or not at all in LMPPs, GMPs and CLPs. GFI1B is important for the development of erythroid progenitors, whereas GFI1 is important for myeloid and T cell development, and the two factors are known to be mutually inhibitory\(^ {37,38} \). Outside the HSC population; Gfi1 was expressed in most LMPPs, CLPs and GMPs, but rarely in PreMegEs, whereas Gfi1b was expressed in most PreMegEs, with lower or absent expression in LMPPs, CLPs and GMPs.

Many genes exhibited heterogeneous expression within cell populations, with some cells expressing the gene at high levels and undetectable expression in others, in line with previous reports of expression heterogeneity in blood stem and progenitor populations\(^ {16–18,22} \). Several transcription factors, including Runx1 and Fli1, had a very similar gene expression distribution in all cell types, and these genes were expressed in almost all analysed cells. Conversely, genes including Erg, Lmo2 and Meis1 differed in expression level between cell types. Genes including Gfi1, Gfi1b and Scl (also known as Tal1) showed bimodal expression amongst the cells that expressed the gene, with the potential therefore to generate three distinct expression states (high, medium, not-expressed) within a single population that is
pure on the basis of FACS analysis. Importantly, such detailed insights into the dynamical nature of transcription factor gene expression in primary blood stem and progenitor cells could not have been obtained from population studies.

**Cell populations can be resolved by differential network activity states**

To establish cell-type-specific patterns of gene expression that may aid our understanding of network activity and cell state transitions, we next performed hierarchical clustering and principal component analysis using the expression data for our transcription factors in all 597 haematopoietic stem/progenitor cells. The relatedness of cells is determined using only the gene expression values, without prior knowledge of which population a cell originates from. Hierarchical clustering demonstrated that messenger RNA levels for these 18 key transcription factors allow the partitioning of cells largely by sorted population (Fig. 3a). This was particularly clear for the GMPs, which formed a distinct cluster. HSCs and PreMegEs formed a cluster separate from the myeloid and lymphoid lineages in which the two populations were also largely separated from each other, whereas LMPPs and CLPs showed significant overlap. There was some mixing of HSCs with LMPPs and PreMegEs, in line with the evidence that LMPPs and the megakaryocyte–erythroid lineage may be generated as early and alternative fates of HSCs (ref. 24) and so are both closely related to HSCs but distinct from one another.

Principal component analysis confirmed the above results, where each data point represents a single cell, colour-coded according to its flow cytometric phenotype (Fig. 3b, upper panel). Principal component 1, which captures the largest proportion of the variation in the data, separates the HSCs and PreMegEs from the lymphoid and myeloid populations, and partially separates the GMPs from the LMPPs and CLPs. Principal component 2 further separates the HSCs and PreMegEs. However, although individual populations can be distinguished there is also significant overlap, particularly between the LMPPs and CLPs, both of which contain myelolymphoid-restricted progenitors99, indicating that cells at the edges of adjacent populations have similar network activity states. *Gata2, Gfi1b, Scl* and *Gfi1* contribute to separation of the HSCs and PreMegEs from the myelolymphoid populations along component 1, whereas *Erg, Hhex* and *Gata1* are important across component 2 (Fig. 3b, lower panel), consistent with known expression patterns in these populations. *Runx1* and *Flt1* contributed little to the separation of cell types, consistent with their similar expression distributions between cell types (Fig. 2).

Gaussian process latent variable models (GPLVMs) are a nonlinear generalization of PCA (ref. 40) and were recently shown to be a powerful alternative, in particular for resolving nonlinear differences in single-cell gene expression patterns41. GPLVM resulted in a better separation of the different cell types than PCA (Fig. 3c), with the greatest improvement found for GMPs. However, LMPPs and CLPs cells could not be resolved completely. Our ability to separate populations was further confirmed by calculating the spatial median, a robust multivariate measure of the centre of the distribution for each cell type, and then analysing distances between cell types. These GPLVM map distances reflected the differentiation hierarchy shown in Fig. 3a with cell types close in the hierarchy located close together in the map (Supplementary Fig. S2).

Whereas for standard PCA the relevance of different genes to the separation of the data (quantified by component loadings in Fig. 3b) can be found only for the entire PCA map, the relevance of each gene can change across the GPLVM map, providing a greater resolution of the changes separating cell types. The GPLVM relevance map (Fig. 3d) shows the most important gene at each point of the map, and illustrates for example that *Gata2, Gfi1* and *Gfi1b* feature frequently throughout the map and particularly in the region bordering the HSC and LMPP populations. Indeed, expression maps for individual genes (Fig. 3e) demonstrate that a high *Gata2* expression level occurs mostly in the HSC and PreMegE populations, whereas high *Gfi1* is restricted mostly to LMPPs, GMPs and a subset of CLPs. Taken together therefore, bioinformatic analysis of single-cell gene expression allowed us to correlate distinct expression states of a core set of 18 key haematopoietic transcription factors with some of the earliest blood stem and progenitor populations, including the earliest known lineage restriction stage from HSCs resulting in distinct megakaryocyte–erythroid and myelolymphoid-restricted pathways.

**Single-cell analysis reveals dynamic regulatory relationships**

We next reasoned that single-cell expression data could be used to identify regulatory linkages by identifying pairs of factors with correlated expression, where a positive correlation suggests that one factor may activate another and a negative correlation indicates an antagonistic relationship. Correlation analysis and hierarchical clustering of the 18 transcription factors across all 597 haematopoietic stem/progenitor cells revealed both positive and negative correlations (Fig. 4a, top left panel). Among the positive correlations is a group of seven genes (*Scl, Gata2, Nfe2, Eto2, Gfi1b, Gata1* and *Ldb1*) known to be important in the megakaryocyte–erythroid lineages35,38,42–49, whereas there was a negative correlation between *PU.1* and *Gata1*, which are thought to function as a switch controlling erythroid and myelomonocytic fate50,51. To establish whether such regulatory relationships were stable or dynamic during differentiation, we repeated the correlation analysis for each of the five stem/progenitor populations individually (Fig. 4a). Although many of the strong positive correlations identified in the whole data set remained stable between cell types, there were some clear differences, particularly in negative correlations, which could suggest that repression, or relief of repression, of some transcription factors by others is a vital step in cell fate transitions. For example, the strong negative correlation between *Gfi1* and *Gata2* present in the whole data set is seen only in HSCs. *Gfi1b* and *Gata1* are negatively correlated with *PU.1*, *Mif*, *Gfi1*, *Ly1* and *Lmo2* in GMPs, and to some extent in CLPs, but are either not correlated or positively correlated in the earlier progenitors and in the megakaryocyte–erythroid precursor. Together these results indicate that expression of the core haematopoietic transcriptional regulatory network is dynamic (Figs 2 and 3), which is presumably intimately connected to the dynamics of regulatory interactions between the components of the network (Fig. 4).

Significant positive and negative correlations between transcription factors were displayed as a putative interaction network (Fig. 4b). Among these is the known *Scl–Gata2* relationship and the *PU.1–Gata1* (ref. 51) and *Gfi1–Gfi1b* (refs 37, 38) inhibitory relationships, indicating...
Figure 3 Single-cell gene expression analysis reveals cell-type-specific regulatory states. (a) Hierarchical clustering of 597 haematopoietic stem and progenitor cells according to the expression of the 18 transcription factors. Coloured bar indicates cell type of origin: Green, HSC; blue, LMPP; purple, CLP; red, GMP; orange, PreMegE. (b) Principal component projections of the 597 haematopoietic stem/progenitor cells, in the first and second components (top), from the expression of all 18 transcription factors. Principal component loadings (bottom) indicate the extent to which each gene contributes to the separation of cells along each component. (c) GPLVM illustrating variations of 18D gene expression patterns between and within cell types in 2D. GPLVMs are a nonlinear generalization of PCA that allow for the analysis of more complex gene expression patterns than PCA and can thus potentially better represent variations between and within populations of different cell types. The uncertainty of the mapping from 2D to the 18D transcription factor space is encoded in grey (white, low uncertainty; grey, high uncertainty). (d) Relevance map showing the most important genes across the GPLVM map. The colours correspond to the distance of the respective gene from the origin on a standard loadings plot (red far away/important, blue close to origin). (e) Expression maps for Gata2 (left) and Gfi1 (right), with high expression in red and low or absent expression in blue. Gata2 is expressed primarily in the HSC and PreMegE clusters, and Gfi1 in LMPPs, GMPs and some CLPs.

that further relationships identified may indeed signify previously unrecognized interactions. Two newly predicted regulatory links (Gata2–Gfi1 and Gata2–Gfi1b) suggested possible involvement of GATA2 in modulating the cross-inhibitory relationship between GFI1 and GFI1B (Fig. 4b), which was of particular interest to us because we had seen downregulation of Gata2 and Gfi1b accompanied by reciprocal upregulation of Gfi1 in a recent transcriptomic and epigenomic analysis of leukaemia development in an MLL-ENL-driven mouse model of acute myeloid leukaemia\(^{32}\) (see Supplementary Fig. S3).
Figure 4 Single-cell expression analysis of haematopoietic transcription factors identifies previously unrecognized putative regulatory interactions between key transcription factors. (a) Hierarchical clustering of Spearman rank correlations between pairs of transcription factors for all 597 cells together and for the different cell types individually as indicated. Genes in all heat maps are ordered according to the clustering performed for all data. Positive correlations (red) may result from the coordinate expression or lack of expression of pairs of factors in individual cells, whereas negative correlations (blue) can result either from the expression of one factor in the absence of the other, or from high expression of one factor and reciprocal low expression of the other in the same cell. (b) Network diagrams showing putative activating relationships between transcription factors suggested by significant positive correlations (top, red lines) and antagonistic relationships suggested by significant negative correlations (bottom, blue lines) in the whole data set. Known relationships are highlighted with bold lines. This highlights a putative transcription factor triad in which Gfi1 is negatively correlated with Gata2 and Gfi1b, but Gata2 and Gfi1b are positively correlated.
Direct repression of Gata2 by GFI1 provides a likely mechanism for negatively correlated expression

To investigate whether downregulation of Gata2 might be mediated directly through GFI1 binding to Gata2 regulatory elements, we interrogated existing chromatin immunoprecipitation (ChIP)-sequencing data for GFI1 in MLL-ENL-transduced cells. Across the Gata2 locus, a prominent peak was identified 83 kilobases (kb) upstream of the Gata2 transcriptional start site (Supplementary Fig. S3B). This −83 kb region had been shown previously to loop to the Gata2 promoter and was bound in the HPC7 haematopoietic progenitor cell line by multiple transcription factors (Supplementary Fig. S3). At present, transcription factor ChIP-seq studies are not possible with the small numbers of cells that can be obtained for the highly purified blood stem cell populations used here for single-cell expression analysis. We nevertheless wanted to confirm binding of GFI1 to the Gata2 gene locus in primary blood cells, and therefore performed GFI1 ChIP-seq in primary mast cells, which like HSCs express the stem cell factor receptor c-KIT and a number of transcription factors important for HSCs including GFI1 and GATA2. This GFI1 ChIP-seq experiment confirmed GFI1 binding to the Gata2 −83 kb region in primary mouse blood cells (Fig. 5a).

As no in vivo activity has as yet been reported for the Gata2 −83 kb region, we generated a LacZ reporter construct with the −83 kb region fused to a minimal SV40 promoter/LacZ reporter cassette. Analysis of LacZ expression in embryonic day (E)11.5 transgenic mouse embryos demonstrated consistent staining in the midbrain, hindbrain and spinal cord (Fig. 5b), all known domains of endogenous Gata2 expression. However, no haematopoietic staining was seen in any of the transgenic embryos. We had shown previously that a Gata2 −3 kb enhancer is active at E11.5 in the dorsal aorta endothelium including budding haematopoietic cells, but not the fetal liver. Given the prominent transcription factor binding to the −83 kb region in haematopoietic cells, we next examined whether a combination of the −83 and −3 kb enhancers was able to drive expression to fetal liver haematopoietic cells. Transgenic embryos carrying a combined enhancer construct (−3/SV/LacZ/−83) exhibited the neural activities of both of the individual enhancers (Fig. 5b). Moreover, staining was seen not only in the dorsal aorta but also in fetal liver haematopoietic cells.

Direct activation of Gfi1b distal enhancer elements by GATA2 provides a likely mechanism for positively correlated expression

To investigate whether GFI1 could repress activity of this element, we next generated a luciferase reporter construct (−3/SV/luc/−83) and performed transfection assays in the HPC7 progenitor cell line, which expresses high levels of Gata2 but low levels of Gfi1. Co-transfection with a Gfi1 expression construct caused a 40% reduction in reporter activity. Luciferase expression is shown relative to −3/SV/luc/−83 and bars are the mean and standard deviation of three biological replicates.

Figure 5 Direct repression of Gata2 by GFI1 through a distal enhancer element provides a mechanism for negatively correlated expression. (a) ChIP-seq analysis of GFI1 in primary mast cells indicates that GFI1 binds to the Gata2 locus at the −83 kb regulatory element. (b) Representative embryos demonstrating LacZ staining for the Gata2 −83, −3 kb and combined −3/−83 kb regulatory element reporter constructs. The −83 kb region alone (SV/LacZ/−83) showed consistent staining of the midbrain, hindbrain and spinal cord, but no haematopoietic staining. The −3 kb enhancer (−3/SV/LacZ) had only hindbrain staining. The −83/−3 kb combined element (−3/SV/LacZ/−83) showed the neural activities of both individual enhancers, but also staining in the dorsal aorta (left-hand panel) and fetal liver haematopoietic cells (right-hand panel). Images of sections taken at ×40 magnification. Scale bar, 50 µm. (c) A luciferase reporter construct carrying both regulatory elements (−3/SV/luc/−83) was transfected into the HPC7 haematopoietic progenitor cell line, which expresses high levels of Gata2 but low levels of Gfi1. Co-transfection with a Gfi1 expression construct caused a 40% reduction in reporter activity.
GATA2, GFI1 and GFI1B suggested by the data. In this regulatory triad, GFI1 and GFI1B are mutually inhibitory, particularly for the promoter as well as the Gfi1b promoter in haematopoietic clusters in the dorsal aorta, and in a subset of fetal liver cells. The +17 kb region shows staining in a small subset of fetal liver haematopoietic cells. Images of sections taken at ×40 magnification. Scale bar, 20 μm. (c) Luciferase reporter constructs carrying the wild-type +16 kb (SV/luc/+16 WT) and +17 kb (SV/luc/+17 WT) regulatory regions transfected in 416B cells exhibited high levels of luciferase activity, particularly for the +16 kb region. Mutation of the two conserved GATA sites and one partially conserved GATA site in the +16 kb region (SV/luc/+16 GATA m12) reduced luciferase activity by >95%. Mutation of the two conserved GATA sites and one partially conserved GATA site in the +17 kb region (SV/luc/+17 GATA m123) also reduced luciferase activity by >95%, m12 and m123 indicate that GATA sites 1, 2 and 3 were mutated. Luciferase activity is shown relative to SV/luc. Experiments were performed in biological duplicate or triplicate on two separate occasions. Shown is one representative experiment with the mean and standard deviation for three biological replicate transfections. (d) A putative regulatory triad including GATA2, GFI1 and GFI1B suggested by the data. In this regulatory triad, GFI1 and GFI1B are mutually inhibitory, whereas GATA2 can activate expression of Gfi1b and GFI1 can repress expression of Gata2.

regions bound by GATA2 in HPC7 as well as a region in the first intron not bound in HPC7 (Fig. 6A). Taken together, transcription factor binding in both mast cells and HPC7 therefore identified 4 candidate regulatory regions that might be involved in mediating control of Gfi1b expression in stem/progenitor cells by GATA2.

To assess whether the four regions bound by GATA2 correspond to bona fide gene regulatory sequences, we performed transgenic assays. LacZ reporter constructs were generated for the Gfi1b promoter as well as the +13, +16 and +17 kb candidate enhancer regions and assayed in E11.5 mouse embryos. The promoter region alone did not mediate any haematopoietic expression, a phenomenon that we have observed before for both Runx1 and Scl/Tal1 promoters. In contrast, all three distal regions mediated haematopoietic expression: the +13 kb region showed weak expression in a subset of circulating (probably primitive) blood cells, the +16 kb region exhibited strong staining in haematopoietic clusters in the dorsal aorta as well as a subset of fetal liver cells, and the +17 kb region showed staining in a small subset of fetal liver haematopoietic cells (Fig. 6B). Transgenic analysis therefore provided in vivo validation of GATA2-bound regions, with the +16 and +17 kb regions being particularly relevant owing to their activity in the anatomical sites of early definitive haematopoietic development.

As transgenic analysis had focused our attention on the +16 and +17 kb enhancer regions as possible mediators of Gfi1b activation by GATA2, we investigated the possible presence of conserved GATA motifs. Both the +16 and +17 kb regions showed extensive sequence conservation across a wide range of mammalian species consistent with their function as gene regulatory elements (Supplementary Fig. S5). Moreover, both enhancers contained two completely conserved GATA motifs. To investigate the role of the GATA sites in enhancer activity, we generated luciferase reporter constructs with both the wild-type GATA and GATA mutant +16 and +17 kb enhancers (Fig. 6c). Following stable transfection assays in the myeloid progenitor cell line 416B, both regions showed substantial enhancer activity that was almost completely lost in the GATA mutant constructs. GATA2 binding in ChIP assays, activity in transgenic assays and presence of conserved GATA motifs essential for enhancer function are therefore all consistent with a model whereby GATA2 directly activates Gfi1b expression through the +16 and +17 kb enhancer regions.

Taken together therefore, single-cell gene expression analysis of primary blood stem and progenitor cells suggests the existence of a regulatory triad including Gata2, Gfi1 and Gfi1b (Fig. 6d), the connectivity of which has been validated using transgenic and transcriptional assays. In this triad, the reported mutual inhibition of GFI1 and GFI1B (refs 37,38) is retained, but is modulated by GATA2 through its activation of Gfi1b and repression by GFI1.

**DISCUSSION**

Cellular phenotypes are controlled by networks of interacting transcription factors. Development can therefore be described as a procession through multiple dynamic regulatory states, which in the case of multilineage differentiation may give rise to multiple distinct outcomes. However, classical networks derived from gene expression or ChIP-seq data provide a population average, giving little insight into cellular heterogeneity and possible regulatory interactions likely to be critical for the lineage commitment/restriction steps of individual progenitor cells.

In this study, we investigated a core transcriptional network of 18 transcription factors in single cells of five related stem and progenitor populations. Bioinformatic analyses were able to broadly distinguish sorted cell populations on the basis only of the expression of those transcription factors, indicating that early haematopoietic stem/progenitor cells are characterized by related but distinct network activity states. Although gross gene expression patterns were consistent with published population studies, we also found significant
heterogeneity in transcription factor expression within populations. This confirms previous observations that FACS-sorted populations are molecularly and functionally heterogeneous at the single-cell level\textsuperscript{16–18}, and suggests either that heterogeneity is an inherent characteristic of stem cells, or that previously unresolved subpopulations may be present that could represent intermediate differentiation steps.

Importantly, heterogeneity did not confound our ability to resolve populations on the basis of their distinct gene expression patterns. Thus, further analysis of defining transcription factor expression patterns and interactions at important lineage restriction stages should facilitate the unravelling of critical transcription factor interactions decisive for lineage commitment steps. Index sorting for example permits the tracing back of each cell to its position in the sorting data, and may therefore allow us in future to link heterogeneous gene expression states to subpopulations. Strikingly, in our study several transcription factors had similar expression levels in all 597 cells analysed, regardless of cell type of origin, including Runx1 and Fli1, whereas others were much more variable. This may suggest that the blood network requires, or is able to tolerate, variation in some factors but not others. Furthermore, correlation analysis revealed both stable and dynamic transcription factor relationships across the cell types analysed, together suggesting that some transcription factor interactions may be vital for the general stability of the network and so remain constant, whereas others are important for network dynamics and state transitions and are therefore more variable.

As transcriptional regulatory networks control the spatiotemporal regulation of lineage-specific genes, successful reconstruction of regulatory hierarchies represents a major step towards gaining a mechanistic understanding of cellular decision-making processes. For example, detailed experimental and computational analysis of a core circuit of Gata2, Scl and Fli1 revealed that this circuit is able to function as a bistable switch, where the internal wiring enables the network to filter noise when responding to external cues\textsuperscript{8,90}. However, the generation of large-scale experimentally validated network models is impeded by the relatively low throughput of experiments that can provide detailed experimental information on the functionality of individual regulatory elements. Network inference, where transcriptional interactions are inferred from statistical dependencies in large expression profiling data sets, provides an alternative approach\textsuperscript{7}. However, in addition to the very substantial costs of performing hundreds of expression profiling experiments, standard microarray or RNA-seq profiling requires large cell numbers, which may not be feasible for rare stem and progenitor populations.

Here we have demonstrated that gene expression data from a large set of single cells present an alternative approach to network reconstruction through the analysis of pairwise correlations between network components. Although hundreds of samples are still required, the fact that these are single cells makes this approach applicable to rare stem cell populations. The robustness of inferred regulatory links was demonstrated by our ability to recover key known regulatory relationships such as the Gata1–PU.1 and Gfi1–Gfi1b antagonisms. Furthermore, we validated two putative regulatory interactions predicted from bioinformatic analysis of our single-cell expression profiling data. This demonstrated that GFI1 directly represses Gata2 expression through the $-83$ kb regulatory region, whereas GATA2 activates Gfi1b through the $+16$ and $+17$ kb regulatory regions. The resultant connectivity resembles a type 2 coherent feedforward loop\textsuperscript{61}, and suggests that GATA2 works within this regulatory triad to modulate the antagonism between GFI1 and GFI1B. Moreover, GATA2 inhibits lymphopoiesis\textsuperscript{62}, and along with Gfi1b is downregulated concurrent with Gfi1 upregulation in the myelolymphoid lineages in our data. This suggests that direct downregulation of Gata2 and Gfi1b by GFI1 may represent a key event during the specification of early lymphoid cells, similar to the role of Tif1γ in modulating the Pu.1–Gata1 antagonism in the erythroid versus myeloid fate choice in zebrafish\textsuperscript{63}.

Loss-of-function mutations in Gata2 have recently been reported to predispose carriers to developing acute myeloid leukaemia\textsuperscript{64}, whereas inhibition of Gfi1 prolonged survival in leukaemia mouse models\textsuperscript{65}. This suggests that identification of network states and reconstruction of network hierarchies from single-cell expression profiling will not only enhance our knowledge of normal differentiation and development, but also provide a blueprint for understanding the subversion of cell fate control likely to underlie many degenerative and malignant pathologies.

### METHODS

Methods and any associated references are available in the online version of the paper.

**Note:** Supplementary Information is available in the online version of the paper.

### ACKNOWLEDGEMENTS

We thank S. A. Clark for help with FACS sorting. Research in the authors’ laboratory is supported by the Medical Research Council, Leukaemia and Lymphoma Research, The Leukaemia and Lymphoma Society, Cancer Research UK, and core support grants by the Wellcome Trust to the Cambridge Institute for Medical Research and Wellcome Trust–MRC Cambridge Stem Cell Institute.

### AUTHOR CONTRIBUTIONS

V.M. designed and performed single-cell experiments, performed analysis and wrote the paper. B.G. conceived the study, designed experiments and wrote the paper. I.C.M. and S.E.J. designed and performed FACS and wrote the paper. F.J.C-N. and A.J. performed preliminary studies. F.J.C-N. performed ChiP-seq experiments. G.S. and M.F.d.B. designed experiments and performed preliminary studies.

### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at www.nature.com/doifinder/10.1038/ncb2709

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

1. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. Cell 132, 631–644 (2008).
2. Ottersbach, K., Smith, A., Wood, A. & Gottgens, B. Ontogeny of haematopoiesis: recent advances and open questions. Br. J. Haematol. 148, 343–355 (2010).
3. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663–676 (2006).
4. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell 51, 987–1000 (1987).
5. Gering, M., Yamada, Y., Rabbitts, T. H. & Patient, R. K. Lmo2 and Scl/Tal1 convert non-axial mesoderm into haemangioblasts which differentiate into endothelial cells in the absence of Gata1. Development 130, 6187–6199 (2003).
6. Di Tullio, A. et al. CCAT1enhancer binding protein alpha (C/EBP(a))-induced transdifferentiation of pre-B cells into macrophages involves no overt retrodifferentiation. Proc. Natl Acad. Sci. USA 108, 17016–17021 (2011).
7. Basio, K. et al. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 37, 382–390 (2005).
et al

Orlic, D., Anderson, S., Biesecker, L. G., Sorrentino, B. P. & Bodine, D. M. 36. Upstream
Yamamoto, M., Takahashi, S., Onodera, K., Muraosa, Y. & Engel, J. D. 34. A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. Bioinformatics 28, 1626–632 (2012).

Shvidasani, R. A. & Orkin, S. H. 37. Erythropoiesis and globin gene expression in mice lacking the transcription factor NF-E2. Proc. Natl Acad. Sci. USA 92, 8690–8694 (1995).

Shvidasani, R. A. et al. 38. Transcription factor NF-E2 is required for platelet formation independent of the actions of thrombopoietin/MGDF in megakaryocyte development. Cell 81, 695–704 (1995).

Hall, M. A. et al. 39. The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S2. Proc. Natl Acad. Sci. USA 100, 992–997 (2003).

Huang, Z. et al. 40. GATA-2 reinforces megakaryocyte development in the absence of GATA-1. Mol. Cell. Biol. 29, 5168–5180 (2009).

Ikonomi, P. et al. 41. Overexpression of GATA-2 inhibits erythroid and promotes megakaryocyte differentiation. Exp. Hematol. 28, 1423–1431 (2000).

Porcher, C. et al. The T cell leukemia oncoprotein SCL/Tal1 is essential for development of all hematopoietic lineages. Cell 86, 47–57 (1996).

Schuh, A. H. et al. ETO-2 associates with SCL in erythroid cells and embryos and represses the expression of globin genes. Blood 121, 1078–1082 (2008).

Grass, J. A. et al. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. Mol. Cell. Biol. 26, 7056–7067 (2006).

Pinto do, O. P., Kolterud, A. & Carlson, L. Expression of the LIM-homoebox gene LHZ generates immortalized steel factor-dependent multipotent hematopoietic precursors. EMBO J. 17, 5744–5756 (1998).

Nardelli, J. et al. Genome-wide analysis of transcriptional reprogramming in mouse models of acute myeloid leukaemia. PLoS One 6, e16330 (2011).

Khandanpour, C. et al. The human GFI136N variant induces epigenetic changes at the Hoxa9 locus and accelerates K-Ras driven myeloproliferative disorder in mice. Blood 120, 4006–4017 (2012).

Grass, J. A. et al. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. Mol. Cell. Biol. 26, 7056–7067 (2006).

Pinto do, O. P., Kolterud, A. & Carlson, L. Expression of the LIM-homoebox gene LHZ generates immortalized steel factor-dependent multipotent hematopoietic precursors. EMBO J. 17, 5744–5756 (1998).

Nardelli, J. et al. Genome-wide analysis of transcriptional reprogramming in mouse models of acute myeloid leukaemia. PLoS One 6, e16330 (2011).

Khandanpour, C. et al. The human GFI136N variant induces epigenetic changes at the Hoxa9 locus and accelerates K-Ras driven myeloproliferative disorder in mice. Blood 120, 4006–4017 (2012).

Grass, J. A. et al. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. Mol. Cell. Biol. 26, 7056–7067 (2006).

Pinto do, O. P., Kolterud, A. & Carlson, L. Expression of the LIM-homoebox gene LHZ generates immortalized steel factor-dependent multipotent hematopoietic precursors. EMBO J. 17, 5744–5756 (1998).

Nardelli, J. et al. Genome-wide analysis of transcriptional reprogramming in mouse models of acute myeloid leukaemia. PLoS One 6, e16330 (2011).

Khandanpour, C. et al. The human GFI136N variant induces epigenetic changes at the Hoxa9 locus and accelerates K-Ras driven myeloproliferative disorder in mice. Blood 120, 4006–4017 (2012).

Grass, J. A. et al. Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. Mol. Cell. Biol. 26, 7056–7067 (2006).
METHODS

Purification of stem cells and progenitor cells. Bone marrow cells were isolated from the bones (femurs, tibiae and crista ilaeca) of 9–12-week-old C57BL/6 mice. For the isolation of LMPs and HSCs, cells were enriched for CD117+ (c-Kit) cells by MACS bead separation with anti-CD117 immunomagnetic beads (Miltenyi Biotec); for the other cell populations, unenriched bone marrow was used. Cells were pre-incubated with Fc-block for the HSC, LMP and CLP stains but not for the myeloid progenitor stain (GMP and Pre-MegE isolation). Cells were stained with antibodies against mouse antigens to allow separation of the individual populations (Supplementary Table S1). A FACS Aria II (BD Biosciences) was used for all cell sorting. Fluorescence-minus-one controls and unstained populations were used as gate-setting controls. Single cells were seeded by an automated cell deposition unit directly into the Fluidigm assay mixture (see below). Test sorts before and after single-cell sorts verified the purity of all populations at >98% on the basis of expression of cell-surface markers (Supplementary Fig. S1). Fluorescent beads were sorted into 96-well plates before and after samples to verify that only single events were sorted into each well.

Single-cell gene expression analysis. Single-cell gene expression analysis was performed using 48.48 Dynamic Array integrated fluidics chips (M48, Fluidigm) on the BioMark HD platform (Fluidigm), which facilitates the simultaneous analysis of 48 genes in each of 48 samples. Complementary DNA synthesis and specific target amplification (preamplification) of genes of interest were performed using the CellsDirect One-Step qRT-PCR kit (Invitrogen). Single cells were sorted by FACS directly into individual wells of 96-well plates containing 5 µl CellsDirect 2 reaction mix (Invitrogen). 0.1 µl SUPERase RNase inhibitor (Ambion), 2.5 µl 0.2 x assay mix, 1.2 µl TE buffer (Invitrogen) and 1.2 µl SuperScripIII Platinum Taq (Invitrogen). The 0.2 x assay mix contained a pool of 24 TaqMan assays (Applied Biosystems; details available on request) at a 1:100 dilution of each assay in TE buffer. Reverse transcription and specific target amplification were performed in the same plates immediately after sorting as follows: 50 °C for 15 min, 95 °C for 2 min, 22 × (95 °C for 15 s and 60 °C for 4 min). cDNA was diluted 1:5 with TE before quantitative PCR (qPCR) on the BioMark HD. cDNA was stored at −20 °C before processing on the BioMark HD. For each population, 6 positive controls of 20 cells per well, 14 negative controls (no cell sorted) and 124 single cells were sorted. This corresponds to 3 M48 Dynamic Arrays per population, each containing 2 positive controls, 4–6 negative controls and 40–42 single cells. For the qPCR, 3 µl of each TaqMan assay was mixed with 3 µl Gene Expression Assay Loading Reagent (Fluidigm). Then, 2.7 µl of diluted cDNA was mixed with 3.2 x TaqMan Universal Mastermix (Applied Biosystems) and 0.3 µl Gene Expression Sample Loading Reagent (Fluidigm). Next, 5 µl of each sample and assay was loaded into individual sample and assay inlets on the M48 Dynamic Array. Samples and assays were then loaded into the reaction chambers of the Dynamic Array using the IFC Controller MX (Fluidigm), and then transferred to the BioMark HD for qPCR (95 °C for 10 min; 40 cycles of 95 °C for 15 s and 60 °C for 60 s).

Testing TaqMan assays. TaqMan assays were tested for single-cell PCR by performing standard curves on the BioMark of cDNA from a population of 100 cells of the HPC7 haematopoietic progenitor cell line, reverse transcribed and preamplified as described above. Assays were selected on the basis of the amplification efficiency and lack of background expression in template controls. The linear range of all assays was used within the sensitivity of the BioMark HD (Ct 7–27).

Bioinformatic analysis of single-cell gene expression data. Single-cell expression data were initially analysed with the Fluidigm Data Collection software. For quality control, amplification curves were quality filtered using a threshold of 0.75 and Ct thresholds were set for each assay, with the same thresholds used across all experiments and cell populations. Data were then exported to Excel as .csv files. All of the Ct values are available in Supplementary Table S3. Samples not expressing any genes (probably as a result of a failure of the sorter to put a cell in the well) were excluded from the analysis (n = 6), as were cells not expressing the housekeeping genes Ubc (n = 0) or Polr2a (n = 17). Expression values over the cutoff of the machine or beyond the linear range of the TaqMan assays (Ct > 27) were set to 28. Each assay was performed in duplicate, and the mean of the duplicates was used for subsequent analysis. Following these quality-control measures, ΔCt values were calculated as previously described28 by cell-wise normalization to the mean expression level of Ubc and Polr2a, as the two most robustly expressed housekeeping genes. Briefly, Ct values were subtracted from the assumed no-template background of the BioMark of 28 (ref. 20), followed by subtraction of the mean Ct value of Ubc and Polr2a for each cell. The ΔCt value for genes that were not expressed was then set to 15, representative of the detection limit of the BioMark. Hierarchical clustering and principal component analysis were performed in R (www.r-project.org). Hierarchical clustering and principal component analysis were performed only on the data for the 18 transcription factors, excluding c-Kit and housekeeping genes. Hierarchical clustering was performed on cells using Spearman rank correlation. Positive and negative correlations between pairs of genes were tested with Spearman rank correlation, with P values calculated on the basis of 10,000 permutations. Positive correlations with a Z score above 12 (P < 3 × 10−5) and negative correlations with a Z score below −4 (P < 6.09 × 10−5) were considered significant, with known antagonistic relationships recovered beyond these values. PCA was performed using the prcomp function.

A nonlinear generalization of PCA, a GPLVM, was employed to generate a nonlinear embedding of the expression data using the FGPLVM toolbox (http://statmath.wu-wien.ac.at/people/N.Lawrence/gfglvm/). A probabilistic mapping was constructed from a two-dimensional (2D) latent space to the 18D data space using Gaussian processes.46 To allow for nonlinear effects we used a radial basis function kernel to construct the covariance matrix. The 2D coordinates of the latent points as well as the kernel parameters were determined by optimizing the likelihood of the data set using 1,000 iterations of a scaled conjugate gradient optimizer48. A gene relevance map—corresponding to a generalized loadings plot for standard PCA—was generated; for this the mapping from latent space to data space was used to calculate the gradient of the expected value for 20 x 20 regularly spaced points across the GPLVM map and the most important gene (greatest norm of gradient) was plotted48. Single-gene expression maps were generated by calculating the expected value of a single gene for 50 x 50 equally spaced points across the GPLVM map. For each cell type the spatial median m in 2D was calculated by minimizing the expected distance between m and the 2D coordinates of all cells of the respective cell type using an extended Weiszfeld algorithm48.

Mouse bone-marrow–derived mast cells. Bone marrow cells were collected from femurs and tibiae of 3- to 5-month-old adult mice. Cells were cultured in Iscove’s modified Dulbecco’s medium (IMDM) supplemented with 10% fetal bovine serum (Sigma), 1% penicillin/streptomycin (Sigma), 150 µM MTG (Sigma), 10% stem cell factor conditioned media from BHK/MKL cells and 10 ng ml−1 of recombinant mIL-3 (Peprotech). Cells were frequently transferred to new flasks to remove adherent cells and experiments were performed after 3 weeks, when cultures were homogeneous. Homogeneity of culture was confirmed by the presence of FcERI by FACS and toluidine blue staining of cytopsins.

ChIP sequencing. ChIP assays were performed as previously described41 using 7 µg polyclonal antibodies against GATA2 (Santa Cruz, sc9008x) and GF1 (Abcam, ab21061) and control nonspecific rabbit IgG (Sigma, I5006). Each sample was amplified2 and sequenced using the Illumina 2G Gene Analyzer, following the manufacturer’s instructions. Sequencing reads were mapped to the mm9 mouse reference genome using Bowtie48, converted to a density plot and displayed as UCSC genome browser custom tracks.

Transgenic mouse analysis and luciferase assays. Luciferase and LacZ reporter constructs were generated using standard recombinant DNA techniques. Coordinates of chromosomal regions cloned are given in Supplementary Table S2. Luciferase assays were performed as described previously8. E11.5 transgenic mouse embryos were generated and LacZ–stained by Cyagen Biosciences. Staining patterns were analysed as described previously31.

Accession numbers. ChIP-seq data for mast cells have been deposited into the NCBI Gene Expression Omnibus portal under the accession number GSE452318.
Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis

Victoria Moignard, Iain C. Macaulay, Gemma Swiers, Florian Buettner, Judith Schütte, Fernando J. Calero-Nieto, Sarah Kinston, Anagha Joshi, Rebecca Hannah, Fabian J. Theis, Sten Eirik Jacobsen, Marella F. de Bruijn and Berthold Göttgens

*Nat. Cell Biol.* **15**, 363–372 (2013); published online 24 March 2013; corrected online 2 April 2013

In the version of this Article originally published, Fig. 6d was incorrect; the red label should have read GFI1. This has now been corrected in the HTML and PDF versions.
Figure S1 Purification of haematopoietic stem and progenitor cells. FACS profiles for the sorting of single (a) CLPs, (b) LMPPs and HSCs, and (c) GMPs and PreMegEs. Also shown are purity tests, verifying the purity of all populations at >98% based on cell surface markers. Each of the 5 progenitor populations specifically studied in this paper have been prospectively purified and functionally characterised at the single cell level in vitro and, in some cases, also in vivo. Importantly, these studies demonstrated that these highly purified populations are enriched for single progenitor cells exhibiting the characteristic lineage outputs that define the overall potential of the cell population. Equally, populations falling outside of the phenotypic definitions applied in these studies exhibit little potential that would argue for their inclusion in the functionally defined populations. As such, each population is highly enriched for cells with specific, and fully or partially restricted lineage potentials, although as for any primary progenitor cell population, individual cells within a phenotypically defined cell population will exhibit a spectrum of both molecular and functional heterogeneity. Never-the-less, any functional variation relevant to the particular lineage restriction step is predicted to occur within the definition of the overall cell population. Moreover, the intrinsic molecular and functional heterogeneity of any phenotypically defined primary progenitor cell population argues the critical importance of a more careful mapping at the single cell level of relevant transcriptional networks that might be shared by a large fraction of individual cells within the cell population.
**Figure S2** Euclidean distances between the spatial medians of the different cell types in the GPLVM map (arbitrary units, as in Figure 3C). Figures in bold correspond to direct links consistent with the typical haematopoietic differentiation hierarchy as shown in Figure 3A. The tree above the table was generated using unsupervised hierarchical clustering, and again reproduces key links from the conventional haematopoietic differentiation hierarchy. Cell types with the smallest distance between their spatial median were CLP and LMPP cells, consistent with their overlapping gene expression profiles.

|       | PreMegE | HSC | LMPP | CLP | GMP |
|-------|---------|-----|------|-----|-----|
| PreMegE | 0       | 0.32| 0.66 | 0.57| 0.84|
| HSC    | 0.32    | 0   | 0.42 | 0.47| 0.72|
| LMPP   | 0.66    | 0.42| 0    | 0.27| 0.38|
| CLP    | 0.57    | 0.47| 0.27 | 0   | 0.27|
| GMP    | 0.84    | 0.72| 0.38 | 0.27| 0   |
Figure S3 Additional ChIP-seq data in MLL-ENL cells and HPC7 haematopoietic progenitor cells shows significant binding of haematopoietic transcription factors at a region 83 kb upstream of Gata2. (a) We recently reported transcriptomic and epigenomic analysis of an MLL-ENL-driven mouse model of AML, in which wild type Lin-Kit+ haematopoietic progenitors, which represent the target cells of leukaemogenic retroviruses, are transformed with a transgene for the MLL-ENL fusion protein. Primary transduced pre-leukaemic cells can be cultured in vitro in the presence of interleukin 3 (IL-3). Microarray analysis demonstrated that Gfi1 is upregulated in this model, with reciprocal downregulation of Gata2 and Gfi1b. (b) Interrogation of ChIP-Seq data in “wild type” HPC7 haematopoietic stem/progenitor cell line and MLL-ENL-transformed cells identified a region 83 kb upstream of the transcriptional start site that was bound by multiple TFs in the HPC7 cells, and GFI1 in MLL-ENL cells. HPC7 cells have been validated as a haematopoietic stem/progenitor cell line and are used in place of primary progenitors which cannot be isolated in sufficient quantities for ChIP analysis.
Figure S4 ChIP-seq data in haematopoietic progenitor cells (HPC7) shows significant binding of GATA2 at multiple Gfi1b regulatory elements, including those bound in mast cells. The GATA2 ChIP-seq data in mast cells from Figure 6A is shown as a comparison.
Figure S5  Multiple species alignments for *Gfi1b* +16 and +17 kb regulatory regions. Shown are the conserved cores of (a) the 952 bp +16 element that runs from mouse chr2:28,454,613-28,455,565, and (b) the 642 bp +17 element that runs from mouse chr2:28,453,418-28,454,060. The mouse sequences including the mutations are labelled ‘Mouse GM’ in blue and were verified by sequencing. Mutated nucleotides are underlined. Nucleotides conserved in all species are highlighted in black, and those conserved in some species are highlighted in grey. GATA binding sites are highlighted in yellow.
Supplementary Table Legends

Supplementary Table 1: Antibodies used for FACS staining.

Supplementary Table 2: Coordinates of genomic regions cloned for *Gata2* and *Gfi1b* regulatory elements. Coordinates are given for the mm9 genome build unless indicated.

Supplementary Table 3: Threshold cycle (Ct) values generated by the BioMark for 24 genes in 620 cells from five haematopoietic stem and progenitor populations. A Ct value of 40 indicates that no amplification was detected. CT values above the limit of detection (Ct 27) have been removed.