An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability

Judith Schütte[1,2], Huange Wang[1], Stella Antoniou[3,*], Andrew Jarratt[3,4,*], Nicola K. Wilson[1], Joey Riepsaame[3], Fernando J. Calero-Nieto[1], Victoria Moignard[1], Silvia Basilico[1], Sarah J. Kinston[1], Rebecca L. Hannah[1], Mun Chiang Chan[3], Sylvia T. Nürnberg[5,6], Willem H. Ouwehand[5], Nicola Bonzanni[7,$], Marella F.T.R. de Bruijn[3,$], Berthold Göttgens[1,$]

[1] Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, United Kingdom and Wellcome Trust - Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, United Kingdom

[2] Present address: Department of Haematology, University Hospital Essen, Essen, Germany

[3] MRC Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 9DS, United Kingdom

[4] Present address: Division of Molecular Medicine, Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Australia

[5] Department of Haematology, University of Cambridge & NHS Blood and Transplant, Cambridge, UK

[6] Present address: University of Pennsylvania, Perelman School of Medicine, Philadelphia PA 19104, USA

[7] IBIVU Centre for Integrative Bioinformatics, VU University Amsterdam, De Boelelaan 1081, Amsterdam, The Netherlands and Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam

* Contributed equally

$ Authors for Correspondence

Abstract

Transcription factor (TF) networks determine cell type identity by establishing and maintaining lineage-specific expression profiles, yet reconstruction of mammalian regulatory network models has been hampered by a lack of comprehensive functional validation of regulatory interactions. Here, we report comprehensive ChIP-Seq, transgenic and reporter gene experimental data that have allowed us to construct an experimentally validated regulatory network model for haematopoietic stem/progenitor cells (HSPCs). Model simulation coupled with subsequent experimental validation using single cell expression profiling revealed potential mechanisms for cell state stabilisation, and also how a leukemogenic TF fusion protein perturbs key HSPC regulators. The approach presented here should help to improve our understanding of both normal physiological and disease processes.

Introduction

Tight regulation of gene expression is essential for both the establishment and maintenance of cellular phenotypes within metazoan organisms. The binding of transcription factor proteins (TFs) to specific DNA sequence motifs represents the primary step of decoding genetic information into specific gene expression patterns. TF binding sites (TFBSs) or motifs are usually short (6-10 bp), and therefore found just by chance throughout the genome. Functional TFBSs often occur as evolutionarily conserved clusters, which in the case of enhancers can act over long distances, thus necessitating comprehensive analysis of entire gene loci to understand the transcriptional control mechanisms acting at mammalian gene loci.

Given the complex regulatory circuitries that arise when control of multiple genes is considered, transcriptional control is often represented in the form of gene regulatory networks (GRNs), which carry most mechanistic information when constructed from detailed knowledge on the TFs and the *cis*-regulatory elements with which they interact (1-6). Importantly, regulatory network models can provide much more than a representation of existing knowledge, because network simulations can reveal possible molecular mechanisms that underlie highly complex biological processes. Boolean modelling approaches have been used to reconstruct core regulatory networks in blood stem cells (7) and myeloid progenitors (8), but neither of these studies took into account the underlying regulatory structure of the relevant gene regulatory elements. Full gene-regulatory information has been used for an ordinary differential equation-based model (9, 10), but was restricted to a small three-gene core circuit. Large consortia efforts such as ImmGen and FANTOM5 have created comprehensive networks of either regulatory elements or gene signatures important for multipotency and differentiation (11, 12). Furthermore, studies looking at gene regulation circuitry in embryonic stem (ES) cells have proposed regulatory networks important for ES cell identity (13, 14). While the complexities of transcriptional control demand approaches such as network modelling, no

single experimental method can provide the complex biological data required for the construction of accurate models. The previously mentioned studies focus their attention on one specific aspect of network modelling and importantly did not combine network analysis with comprehensive functional validation. Given that the key building blocks are gene regulatory sequences and the TFs bound to them, essential information for network reconstruction includes (i) comprehensive TF binding data, (ii) *in vivo* validation of the functionality of TF-bound regions as bona fide regulatory elements, and (iii) molecular data on the functional consequences of specific TF binding events (e.g. activation vs. repression). The regulatory network model that we present in this study comprises all of the aforementioned components and is accompanied by functional validation of model predictions.

Results

In vivo validation of *cis*-elements as regulatory network nodes connecting 9 HSPC TFs

For the reconstruction of a core GRN model for HSPCs, we focussed on nine major HSPC regulators (ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1, TAL1), for which genome-wide binding patterns in the murine multipotent progenitor cell line HPC7 have previously been published (15). First, we searched the literature to summarise known *cis*-regulatory regions for the nine TFs that possess haematopoietic activity in transgenic mouse embryos, which recovered a total of 14 regions: *Erg*+85 (16), *Fli1*-15 (17), *Fli1*+12, *Gata2*-3 (18), *Gata2*+3 (= *Gata2*+9.5) (19), *Gfi1b*+13, *Gfi1b*+16, *Gfi1b*+17 (16, 20), *Lyl1* promoter (21), *Spi1*-14 (22), *Runx1*+23 (23), *Tal1*-4 (24), *Tal1*+19 (25) and *Tal1*+40 (26).

To extend this partial knowledge of relevant gene regulatory sequences to a comprehensive definition of how these nine TFs might cross-regulate each other, we made use of the genome-wide binding data for the nine TFs (15) as well as information on acetylation of histone H3 at

lysine 27 (H3K27ac) (27) in the HPC7 blood progenitor cell line. Additional candidate gene regulatory regions for all nine TFs were selected based on binding of at least three TFs and H3K27ac, since it has been shown previously that transcriptionally active regions are commonly bound by multiple TFs and display H3K27 acetylation (28). To assign putative candidate regions to a given TF, they had to be located between its respective upstream and downstream flanking genes, i.e. within the gene body itself or its 5' and 3' intergenic flanking regions. The *Erg* gene locus for example contains five candidate *cis*-regulatory regions based on these criteria, namely *Erg*+65, *Erg*+75, *Erg*+85, *Erg*+90 and *Erg*+149 (Fig. 1a), of which only the *Erg*+85 region had previously been tested in transgenic mice (16). Inspection of the gene loci of all nine TFs resulted in the identification of 35 candidate *cis*-regulatory regions (Fig. 1b, Fig. 1-figure supplements 1-8). In addition to the 14 haematopoietic enhancers previously published, eight of the 35 new candidate regulatory regions had previously been shown not to possess activity in tissues of the blood system of mouse embryos: *Gata2*-83 (*Gata2*-77), *Gfi1b* promoter (20), *Spi1*-18, *Spi1* promoter (22), *Runx1* P1 promoter (29), *Tal1*-9, *Tal1* promoter (30) and *Tal1*+6 (31). Of the remaining 27 candidate *cis*-regulatory regions, two coincided with genomic repeat regions (*Runx1*-322 and *Runx1*+1) and were excluded from further analysis because mapping of ChIP-Seq reads to such regions is ambiguous. Since a comprehensive understanding of regulatory interactions among the nine HSPC TFs requires *in vivo* validation of candidate regulatory regions, we next tested the remaining 25 candidate *cis*-regulatory regions for their ability to mediate reporter gene expression in embryonic sites of definitive haematopoietic cell emergence and colonisation, namely the dorsal aorta and foetal liver of E10.5 to E11.5 transgenic *LacZ*-reporter mouse embryos. For the *Erg* locus, this analysis revealed that in addition to the previously known *Erg*+85 enhancer, the *Erg*+65 and *Erg*+75 regions also displayed activity in the dorsal aorta and/or the foetal liver, while the *Erg*+90 and *Erg*+149 regions did not (Fig. 1c). Careful

inspection of a total of 188 transgenic mouse embryos revealed that nine of the 25 identified regions showed *LacZ* expression in the dorsal aorta and/or foetal liver (Fig. 1b, Fig. 1-figure supplements 1-9). This large scale transient transgenic screen therefore almost doubled the number of known *in vivo* validated early haematopoietic regulatory elements for HSPC TFs.

ChIP-Seq maps for a second progenitor cell line validate core regulatory interactions

Although HPC7 cells are a useful model cell line for the prediction of genomic regions with haematopoietic activity in transgenic mouse assays (16), they are refractory to most gene transfer methods and therefore not suitable for functional characterisation of regulatory elements using standard transcriptional assays. By contrast, the 416b myeloid progenitor cell line can be readily transduced by electroporation and therefore represents a candidate cell line for functional dissection of individual regulatory elements. As ChIP-Seq profiles in 416b cells had not been reported previously, we performed ChIP-Seq for H3K27ac and the nine TFs in this cell line (Fig. 2a, Fig. 2-figure supplements 1-8). Alongside with our previously published HPC7 data, this new 416b dataset represents the most complete genome-scale TF-binding analysis in haematopoietic progenitor cell lines to date, with all new data being freely accessible under the following GEO accession number GSE69776 and also at http://codex.stemcells.cam.ac.uk/. Genome-wide TF binding patterns in 416b and HPC7 cells were closely related when compared with binding profiles for the same factors in other haematopoietic lineages (Fig. 2b, Fig. 2-figure supplement 9). Inspection of the gene loci for the nine HSPC TFs not only revealed many similarities between 416b and HPC7 cells, but also some differences in TF binding patterns. Overall, TF occupancy at the 23 regions with activity in haematopoietic tissues (14 previously published (16-26) and 9 newly identified) does not change between the two cell types in 71 % of all cases (147 of 207 binding events), is gained in

416b cells in 16 % (33 of 207) and lost in 13 % (27 of 207) of cases compared to HPC7 cells (Fig. 2c). Next, all 23 elements were filtered to only retain those elements which were bound by at least 3 of the 9 TFs and displayed elevated H3K27ac in HPC7 and 416 cells. This led to the removal of the *Gata2*-3, which is not bound by any of the nine TFs in either cell type, *Gata2*-92 and *Gfi1b*+13, which are only bound by one or no TFs in 416b cells, and *Fli1*-15, which is not acetylated in 416b cells (Fig. 2c, Fig. 2-figure supplements 1-3). Overall, 19 *cis*-regulatory regions were therefore taken forward as a comprehensively validated set of regions for the reconstruction of an HSPC regulatory network model.

Comprehensive TFBS mutagenesis reveals enhancer-dependent effects of TF binding

The reconstruction of a core regulatory network model requires information about the effect of TF binding on gene expression, which can be activating, repressing or non-functional. In order to analyse the effects of all TF binding events at all 19 regulatory regions, we performed luciferase reporter assays in stably transfected 416b cells. Based on multiple species alignments between five species (mouse, human, dog, platypus, opossum), we identified conserved TFBSs for the nine TFs (Fig. 3a, Fig. 3-figure supplements 1-19), and generated mutant constructs for each of the 19 regulatory regions, resulting in 87 reporter constructs that were tested by luciferase assays (19 wild-type, 68 mutants). To ensure that DNA binding of the TFs was abrogated, the key DNA bases involved in DNA-protein interactions were mutated and the resulting sequences were scanned to ensure that no new binding sites were created (32). For each of the 19 regulatory regions the conserved TFBSs were mutated by family, for example all six Ets sites within the *Erg*+65 region were mutated simultaneously in one construct and this element was then treated as the *Erg*+65_Ets mutant. TFBS mutations reduced or increased activity compared to the wild-type enhancer, or indeed had no significant

effect (Fig. 3b, Fig. 3-figure supplements 1-18). For instance, at the *Erg*+65 region, mutation of the six Ets binding sites or the three Gata binding sites reduced luciferase activity, whereas mutation of the three Ebox or the three Gfi motifs increased luciferase activity (Fig. 3b). Comparison of the luciferase assay results for all 19 *cis*-regulatory regions (Fig. 3c) reveals that for each motif class mutation can result in activation, repression or no-change. This observation even extends to single gene loci, where for example mutation of the Gata site reduced activity of the *Erg*+65 region, but increased activity of the *Erg*+85 enhancer (Fig. 3c). Taken together, this comprehensive mutagenesis screen highlights the dangers associated with extrapolating TF function simply from ChIP-Seq binding events and thus underlines the importance of functional studies for regulatory network reconstruction.

Dynamic Bayesian network modelling can incorporate complex regulatory information and shows stabilization of the HSPC expression state

We next set out to construct a regulatory network model that incorporates the detailed regulatory information obtained for potential cross-regulation of the nine HSPC TFs obtained in the previous sections (summarised in Fig. 4a). We focussed on three categories of causal relationships: (i) one or several TFs can bind to a certain type of motif at a given regulatory region, and the probability of a motif being bound depends on the expression levels of the relevant TFs; (ii) TFBS mutations at a given regulatory region altered luciferase activities compared to the wild-type, thus capturing the impact of TF binding on activity of the given regulatory region; (iii) individual regulatory regions show varying degrees of activation over baseline controls, which translate into different relative strengths of individual *cis*-regulatory regions. To incorporate this multi-layered experimental information, we constructed a three-tier dynamic Bayesian network (DBN) to jointly represent all those causal relationships (see

Material and Methods and Fig. 4b). The reconstructed DBN represents a first-order time-homogeneous Markov process, which is a stochastic process where the transition functions are the same throughout all time points and the conditional probability distribution of future states depends only on the present state (see Material and Methods). The model is calculated so that the expression at t+1 is influenced by the expression at t0; analogously, the expression at t0 is influenced by the expression at t-1, and so on. Therefore, though the model does not incorporate "epigenetic memory", past expression levels directly influence current expression levels. Model execution therefore permits the simulation of gene expression states in single cells over time, as well as the calculation of gene expression distributions for each gene across a population of simulated single cells.

Having generated a DBN model incorporating extensive experimental information, we next investigated the expression states following model execution. First, we investigated whether the network model was compatible with the HSPC expression profile from which all the experimental data are derived, namely co-expression of all nine TFs. To this end, model execution was initiated with expression levels for all nine TFs set at the midpoint level of 0.5. A representative single cell modelled over time rapidly adopts characteristic levels of expression for each of the nine genes, with some genes showing perpetual fluctuations (Fig. 4c). The same expression levels were reached when the model was initiated with expression starting at 0.2, 0.8 or with initially only FLI1, RUNX1 and TAL1 being expressed at 0.5 (Fig. 4-figure supplement 1). We next modelled the overall distribution of the nine TFs as might be seen in a cell population by running 1000 model simulations (Fig. 4d). This analysis demonstrated that our model is compatible with co-expression of all nine genes within the same single cell. Moreover, stable expression over time for some genes as well as oscillations around a characteristic mean expression level for other genes suggests that our model may have

captured those aspects of HSPC regulatory networks that ensure maintenance of stem/progenitor cells.

Relative stability to experimental perturbation is recapitulated by the model

The TFs TAL1 and LYL1 are important regulators of adult haematopoiesis, but the deletion of each factor individually has only minor effects on adult HSC function (33-35). Combined deletion in adult HSCs however causes a severe phenotype with rapid loss of HSPCs (36). We wanted to investigate to what extent our computer model could recapitulate these known phenotypes through *in silico* perturbation simulations. To quantify if a change in the expression profile of a given TF was significant we performed a Wilcoxon rank-sum test. Interestingly, this significance calculation demonstrated that both large and small fold-changes can be significant. Simulated perturbation of just LYL1 caused significant alterations to the expression profiles of *Gfi1b*, *Tal1*, *Fli1* and *Gata2*, but none of these were associated with a substantial shift in mean expression levels (Fig. 5a, Fig. 5-figure supplement 1). Perturbation of just TAL1 caused significant changes to the expression profiles of *Runx1*, *Gfi1b* and *Gata2*, and again none of these were associated with a substantial shift in expression levels (Fig. 5b, Fig. 5-figure supplement 1). Simultaneous deletion of both factors caused significant changes in gene expression profiles in all TFs except for *Fli1*. Unlike for the single TF perturbations, *Gata2* and *Runx1* showed substantial shifts in expression levels when both LYL1 and TAL1 were simulated to be knocked down (Fig. 5c, Fig. 5-figure supplement 1). Of note, the significance calculations highlight that there may be no one perfect way to visualize these small fold-change alterations. We therefore also generated histogram plots as an alternative visualization (Fig. 5-figure supplement 2).

229 We next wanted to compare model predictions with actual experimental data in the 416b cell
230 line, from which the information for model construction had been derived. Because our DBN
231 model is particularly suited to model the expression states in single cells, we compared
232 predicted and experimentally observed effects of knockdown or overexpression in single cells.
233 To this end we knocked down the expression of TAL1 in 416b cells by transfecting the cells
234 with siRNA against *Tal1* (siTal1) or control siRNA (siCtrl). Forty-eight hours after
235 transfection, gene expression for the nine network genes was analysed in 44 siTal1 treated cells
236 and 41 siCtrl treated cells. Importantly, 29 of 44 cells (66%) transfected with siTal1 showed no
237 expression of *Tal1* anymore, demonstrating the successful knockdown (Fig. 5d, Fig. 5-source
238 data). Down-regulation of TAL1 caused a significant change in the expression profiles of
239 *Tal1*, *Fli1* and *Gfi1b*, but a substantial shift of median expression was only observed for *Tal1*
240 (Fig. 5-figure supplement 1). Experimental validation therefore confirmed the occurrence of
241 statistically significant small-fold changes in expression profiles following single TF
242 knockdown, although there was no perfect match between the genes affected in the model and
243 experiment. To extend comparisons between model predictions and experimental validation,
244 we investigated the consequences of knocking down the expression of PU.1 and
245 overexpressing GFI1B. Complete removal of PU.1 *in silico* after the model had reached its
246 initial steady state had no effect on the expression levels of the other TFs (Fig. 6a). To
247 investigate whether the model prediction is comparable to experimental data obtained from
248 single cells, single cell gene expression analysis using the Fluidigm Biomark HD platform was
249 performed using 416b cells transduced with shRNA against PU.1 (shPU.1) or luciferase
250 (shluc). Three days after transduction, 121 shPU.1 and 123 shluc transduced single cells were
251 analysed for their expression of *Spi1* and the other eight TFs of the network. 18 shPU.1-
252 transduced cells (15%) showed a complete loss of *Spi1*, and expression of *Spi1* in the
253 remaining cells was markedly reduced compared to the control cells (shluc) (Fig. 6a, Fig. 5-

source data), highlighting the efficiency of the PU.1 knockdown. *Spi1*, *Runx1*, *Erg* and *Fli1* showed a significant change in expression profiles after the depletion of PU.1, but this involved a substantial shift in median expression levels only for *Spi1* and *Runx1* (Fig. 5-figure supplement 1). Expression profiles of the remaining five TFs did not change as a result of reduced PU.1 levels (Fig. 6a, Fig. 5-source data), therefore mostly confirming the model prediction.

Next, we modelled GFI1B overexpression *in silico* by increasing the expression level of *Gfi1b* to the maximum value after the model had reached its initial steady state which led to a significant change in the expression profiles of *Gfi1b*, *Meis1*, *Erg* and *Runx1*, although only *Gfi1b* and *Meis1* showed a substantial shift in median expression levels (Fig. 6b, Fig. 5-figure supplement 1, Fig. 5-source data). Expression profiles of the other five TFs were unaltered. Single cell gene expression analysis of 90 single 416b cells transduced with a Gfi1b-expressing vector and 104 single 416b cells transduced with an empty control vector showed a significant increase in the expression of *Gfi1b* and a significant alteration to the expression profile of *Erg*, but only the changes to *Gfi1b* involved a substantial shift in median expression levels. No significant expression changes were seen in any of the other seven network genes (Fig. 6b). Both PU.1 and GFI1B perturbation studies therefore emphasize the resilience of the HSPC TF network to single TF perturbation. Moreover, our *in silico* model reflects this, thus suggesting that the comprehensive experimental information used to construct the network model has allowed us to capture key mechanistic aspects of HSPC regulation. Of note, there were no short-term major expression changes immediately after the perturbation in the silico simulations for the three single TF perturbation described above. For completeness we performed *in silico* modelling for all permutations of single TF knockdown / overexpression as well as all pairwise combinations of all 9 TFs analysed (a total of 162 simulations, Fig. 6-figure supplement 1).

Major perturbations by the AML-ETO oncoprotein are captured by the network model

As the TF network described above is relatively stable to single TF perturbations, we set out to test whether a simulation that mimics the situation present in leukemic cells can influence the expression states of the nine TFs in our network. The *Aml-Eto9a* translocation is amongst the most frequent mutations in AML (reviewed in (37)). The resulting fusion protein is thought to act in a dominant negative manner to repress RUNX1 target genes. To simulate the leukemic scenario caused by AML-ETO expression, we fixed the level of *Runx1* to be the maximum value 1 and at the same time converted all activating inputs of RUNX1 to inhibiting inputs in our DBN model. Interestingly, this simulation of a "leukemic" perturbation caused significant expression changes to all eight of the core HSPC TFs (Fig. 6c). To compare the AML-ETO simulation results with experimental data, we utilised a doxycycline-inducible expression system to generate 416b cells with inducible expression of AML1-ETO fused to a mCherry reporter via a self-cleaving 2A peptide spacer. Following doxycycline induction, 56 single mCherry positive and 122 single mCherry negative 416b cells were analysed by single cell gene expression. Significant gene expression changes can be seen in six of the nine core HSPC TFs (all except *Tal1*, *Erg* and *Gata2*) thus highlighting significant overlap between predictions and experimental validation, although there are also notable differences between model predictions and the experimental data (see for example Gata2; Fig. 6c, Fig. 5-figure supplement 1, Fig. 5-source data). These results demonstrate that our new HSPC network model can capture many gene expression changes caused by ectopic expression of a leukemia oncogene as well as providing a useful model for normal HSPC transcriptional regulation. The inability of any model to completely recapitulate experimental data is not unexpected. Possible reasons in our case may include more complex activities of the onco-fusion protein than would be captured by our assumption that its "only" function is as a straightforward dominant-negative

effect, or the fact that the computational model is a closed system of only the 9 network TFs, whereas the experimental single cell perturbation is subject to possible knock-on consequences from gene changes outside of the 9-TF network.

Discussion

Transcription factor networks are widely recognised as key determinants of cell type identity. Since the functionality of such regulatory networks is ultimately encoded in the genome, the logic that governs interactions between network components should be identifiable, and in due course allow for the construction of network models that are capable of capturing the behaviour of complex biological processes. However, the construction of such network models has so far been severely restricted because the identification and subsequent functional characterisation of mammalian regulatory sequences represent major challenges, and the connectivity and interaction rules within regulatory networks can be highly complex. Here we report a comprehensive mammalian transcriptional network model that is fully grounded in experimental data. Model simulation coupled with subsequent experimental validation using sophisticated single cell transcriptional assays revealed the mechanistic basis for cell state stability within a haematopoietic progenitor model cell line, and also how a leukemogenic TF fusion protein can perturb the expression of a subset of key blood stem cell regulators.

Pictorial representations of putative network models are commonly shown in publications reporting ChIP-Seq TF binding datasets (38). However, due to the lack of experimental underpinning, such representations are simple images that do not encode any of the underlying gene regulatory logic, and importantly therefore cannot provide executable computational models that can be used to simulate biological systems. Although the experimentally-grounded network model shows good agreement with the relative expression states of the nine TFs for

328 the wild-type as well as the perturbation data, model predictions are not correct in all cases.
329 Apart from the obvious caveat that any computer model is an abstraction of reality and
330 therefore will not be correct in every detail, it also needs to be stressed that we treat the nine
331 TFs as an isolated module for the computer simulations, and therefore could not account for
332 possible influences by additional genes that may affect single cell gene expression
333 measurements in the perturbation experiments.

334 Statistical significance calculations demonstrated that both the computer model and the
335 experimental data showed significant changes in gene expression profiles that were associated
336 with minimal fold-change alterations to median expression levels. Such alterations to
337 expression profiles were prevalent in both single and double-gene perturbations, whereas
338 substantial shifts in median expression were mostly restricted to the double perturbations (and
339 also the AML-ETO oncogene overexpression). This observation suggests that (i) our approach
340 has the capacity to reveal aspects of the fine-grained nature of biological networks, and (ii) the
341 network presented in this study is largely resistant to perturbations of individual TFs in terms
342 of substantial fold-change alterations in median expression levels. We believe that it may well
343 be possible that the statistically significant small-fold changes in HSPC network genes may be
344 responsible for the mild phenotypes seen when major HSPC regulators are deleted in adult
345 HSPCs. *Tal1*$^{-/-}$ mice for example are not viable because TAL1 is absolutely required for
346 embryonic blood development (39), yet deletion of TAL1 in adult HSCs only causes minor
347 phenotypes (33). Another noteworthy observation is that it would have been impossible to
348 detect the statistically significant yet small fold-changes using conventional expression
349 profiling, because they only become apparent following the statistical analysis of expression
350 distributions generated by assaying lots of single cells. More generally it is important to
351 acknowledge that the question of how close the present model comes to capturing the

underlying biological processes can only be revealed through much more exhaustive experimental validation studies.

A potential caveat for network reconstruction based on identification of regulatory elements comes from the difficulties associated with capturing negative regulatory elements. As shown elegantly for CD4 and CD8 gene silencers in the lymphoid lineage, TFs involved in the early repression of a locus are not required for maintenance of the silenced state (40, 41). Identification of negative regulatory inputs may therefore require an expansion of datasets to look across sequential developmental stages. It will therefore be important in the future to extend the work presented here to include additional HSPC regulators as well as additional stages along the haematopoietic differentiation hierarchy. Of note, TF-mediated cellular programming experiments have demonstrated that modules of 3-4 TFs are able to confer cell-type specific transcriptional programmes (42-45), consistent with the notion that a network composed of nine key HSPC regulators is able to capture useful information about HSPC regulatory programmes.

One of the most striking observations of the regulatory network defined here is the high degree to which the HSPC expression state is stabilised. As such, this model is different from previous experimentally-grounded transcriptional regulatory network models (46). These earlier model organism networks have inherent forward momentum, where the model captures the progression through successive embryonic developmental stages characterised by distinct expression states.

The model reported here is based on and validated with data from haematopoietic progenitor cell lines, which can differentiate (47, 48), but can also be maintained in stable self-renewing conditions. A recent study by Busch and colleagues tracked labelled $Tie2^{+}$ HSCs in the bone marrow, and showed that haematopoietic progenitors *in vivo* are also characterised by a

substantial self-renewal capability, therefore highlighting the stable state in which they can reside for several months (49). The observed stability of the HSPC expression state presented here is therefore likely to capture aspects of the regulatory mechanisms maintaining the steady state of primary haematopoietic progenitor cells, a notion reinforced further by the fact that our model is based on *in vivo* validated regulatory elements.

The two types of models therefore accurately capture the properties of the distinct biological processes, e.g. driving developmental progression on the one hand, and maintaining a given cellular state on the other. Different design principles are likely to be at play, with feed-forward loops representing key building blocks of early developmental GRNs, while the network described here shows an abundance of auto-regulatory feedback loops and partially redundant enhancer elements, both of which may serve to stabilise a given cellular state.

Of particular interest may be the organisation of the *Runx1* gene locus, where RUNX1 protein provides positive feedback at some, and negative feedback at other HSPC enhancers. Given that these different enhancers employ overlapping yet distinct sets of upstream regulators, it is tempting to speculate that such an arrangement not only stabilises a given expression level, but also provides the means to either up- or down-regulate RUNX1 expression in response to diverse external stimuli that may act on specific RUNX1 co-factors at either the repressing or activating RUNX1 binding events. Taken together, we report widely applicable experimental and computational strategies for generating fully validated regulatory network models in complex mammalian systems. We furthermore demonstrate how such a model derived for blood stem/progenitor cells reveals mechanisms for stabilisation of the progenitor cell state, and can be utilised to analyse core network perturbations caused by leukemic oncogenes.

Materials and Methods

ChIP-Sequencing and data processing

The mouse myeloid progenitor 416b cell line (48) was received from Chester Beatty lab and confirmed to be mycoplasma free. The cells were cultured in RPMI with 10 % FCS and 1 % Penicillin/Streptomycin.

ChIP assays were performed as previously described (16, 27), amplified using the Illumina TruSeq ChIP Sample Prep Kit and sequenced using the Illumina HiSeq 2500 System following the manufacturer's instructions. Sequencing reads were mapped to the mm10 mouse reference genome using Bowtie2 (50), converted to a density plot and displayed as UCSC genome browser custom tracks. Peaks were called using MACS2 software (51). Mapped reads were converted to density plots and displayed as UCSC genome browser custom tracks. The raw and processed ChIP-Seq data have been submitted to the NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) and assigned the identifier GSE69776. A binary binding matrix was created using in-house scripts, clustered using the dice coefficient and a heatmap was plotted using gplots in R in order to compare newly generated ChIP-Seq data with previously published data (52).

Analysis of enhancer activity in transient transgenic mouse embryos

Genomic fragments spanning the candidate *cis*-regulatory regions were generated by PCR or ordered as gBlocks (Life Technologies GmbH) and cloned downstream of the LacZ gene in an hsp68LacZ (Runx1 constructs) or SVLacZ (all other constructs) reporter vector. Coordinates of candidate chromosomal regions and corresponding primer sequences are given in Fig. 3-figure supplement 20. For Runx1, E10 mouse transient transgenic embryos carrying LacZ enhancer-reporter constructs were generated by pronuclear injection of (C57BL/6 x CBA)/F2 zygotes following standard procedures. Transgenic embryos were identified by LacZ-specific PCR on

genomic DNA isolated from yolk sac (5'-GCAGATGCACGGTTACGATG-3'; 5'-GTGGCAACATGGAAATCGCTG-3'). Xgal staining and cryostat sectioning were performed as previously described (23). Embryos were photographed using a Leica MZFLIII microscope, Leica DFC 300F digital camera (Leica Microsystems, Milton Keynes, UK) and Openlab software (Improvision, Coventry, UK) and sections were examined using a Nikon Eclipse E600 microscope (Nikon, Japan) equipped with 20x and 40x Nomarski objectives. Photographs were taken using a Nikon DXM 1200c Digital Camera (Nikon, Tokyo, Japan). E11.5 transient transgenic embryos of all other candidate *cis*-regulatory regions were generated by Cyagen Biosciences Inc (Guangzhou, China). Whole-mount embryos were stained with 5-bromo-4-chloro-3-indolyl-β-d-galactopyranoside (X-Gal) for β-galactosidase expression and photographed using a Nikon Digital Sight DS-FL1 camera attached to a Nikon SM7800 microscope (Nikon, Kingston-upon-Thames, UK). Candidate transgenic mouse embryos with LacZ staining in haematopoietic tissues were subsequently embedded in paraffin, stained with 0.1 % (w/v) Neutral Red and cut into 6 μm deep longitudinal sections. Images of sections were acquired with a Pixera Penguin 600CL camera attached to an Olympus BX51 microscope. All images were processed using Adobe Photoshop (Adobe systems Europe, Uxbridge, United Kingdom).

Luciferase reporter assays

Wild-type and mutant DNA fragments for candidate regulatory regions were either cloned using standard recombinant DNA techniques, ordered as gBlocks (Life Technologies) or obtained from GeneArt® by Life Technolgies. DNA fragments were cloned into pGL2 basic or pGL2 promoter vectors from Promega using restriction enzymes or by Gibson Assembly. TFBSs for the nine TFs of interest (corresponding DNA sequences are listed in Fig. 3-figure supplement 19) were identified based on multiple species alignments between five species

449 (mouse, human, dog, platypus, opossum). Where a region contained multiple instances of the
450 same motif, a single mutant construct with all relevant motifs mutated simultaneously was
451 generated (for generated point mutations check Fig. 3a and Fig. 3-figure supplements 1-18).
452 Where TF binding was observed in ChIP-Seq experiments in 416b cells, but the TFBS was not
453 conserved, the motifs present in the mouse sequence were mutated. Stable transfections of the
454 416b cell line were performed using 10 µg reporter construct, 2 µg neomycin resistance
455 plasmid and $1x10^7$ 416b cells in 180 µl culture medium per pulse. The sample was
456 electroporated at 220V and capacitance of 900 µF using the GenePulser Xcell Electroporation
457 System (Bio-Rad). Immediately after transfection, the sample was split into four culture plates.
458 Twenty-four hours after transfection Geneticin G418 (Gibco by Life Technologies) at a final
459 concentration of 0.75 mg/ml was applied to the culture to select for transfected cells. The
460 activity of the luciferase reporter constructs was measured 12-16 days after transfection by
461 using a FLUOstar OPTIMA luminometer (BMG LABTECH). The luciferase activity was
462 normalised to the cell number and presented as relative activity compared to the wild-type
463 construct. All assays were performed at least three times in quadruplicates.

464

465 **Single cell gene expression and data analysis**

466 The TAL1 knockdown was performed using pools of siRNA against Tal1 (Dharmacon) which
467 were transfected into 416b cells. Briefly, 1 x 10^6 cells were electroporated with either a control
468 or Tal1 siRNA. Forty-eight hours after transfection, cells were sorted into 96 well PCR plates
469 containing lysis buffer using the BD Influx Cell Sorter.

470 The PU.1 knockdown was performed as previously described (27).

471 The MigR1-Gfi1b retroviral expression vector and the corresponding empty vector control (53)
472 were used for GFI1B overexpression. Two million 416b cells were transduced with the above

473 listed vectors by adding viral supernatant and 4 μg/ml polybrene to the cells, followed by
474 centrifugation at 900 x g for 90 min at 32ºC and incubation with 5% CO_2 at 32ºC. Half of the
475 media was then replaced with fresh culture media and cells were incubated at 37ºC with 5 %
476 CO_2. Forty-eight hours after transduction, GFP^+ cells for each cell population were sorted into
477 96 well PCR plates containing lysis buffer using the BD Influx Cell Sorter.

478 To induce AML1-ETO9a expression, the 416b cell line was co-transfected with: 1) a plasmid
479 containing the tetracycline transcription silencer (tTS), the tetracycline transactivator (rtTA)
480 and blasticidine resistance under the control of the *EF1α* promoter; 2) a plasmid containing the
481 entire *Aml-Eto9a* cDNA (obtained from vector MigR1-AE9a, Addgene no. 12433) in frame
482 with a F2A element and the mCherry protein under the control of a tetracycline responsive
483 element; and 3) transposase PL623 (54) (kindly donated by Pentao Liu, Sanger Institute,
484 Cambridge) in order to promote simultaneous stable integration of the two constructs described
485 above. After 6 days of culture without selection, cells were incubated with 1 μg/ml of
486 Doxycycline for 24 hours and then stained with DAPI. mCherry positive and negative cells that
487 did not stain with DAPI were sorted into 96 well PCR plates containing lysis buffer using the
488 BD Influx Cell Sorter.

489 Single cell gene expression analysis was performed using the Fluidigm BioMark platform
490 followed by bioinformatics analysis as previously described (20). All cells that express less
491 than 48 % of genes assayed were removed from the analysis for PU.1 knockdown and GFI1B
492 over-expression, all cells expressing less than 56 % of genes assayed were removed from the
493 TAL1 knockdown and all cells that express less than 44 % of genes assayed were removed
494 from the analysis for the AML-ETO9a induction. Importantly, this thresholding resulted in the
495 removal of similar numbers of cells in both the perturbation and control arms of the
496 experiments. The raw data as well as the normalised data (normalised to Ubc and Polr2a) of the
497 gene expression analysis are listed in Fig. 5-source data).

Computational modelling

The first-order DBN shown in Fig. 4b was established on the basis of regulatory information summarized in Fig. 4a. The DBN essentially presents a discrete-time stochastic process that has the Markov property, i.e. the state of the process at the next time point depends purely on its state at the current time point. Also note that this is a time-homogeneous (or time-invariant) DBN, where the transition functions/matrices are the same throughout all time points.

To specify parameters of the DBN, we defined a motif family at a specific regulatory region as a unique binary variable; with value "1" indicating that no motif of a motif family is bound at the specific region and value "2" indicating that at least one motif of the motif family is bound by a TF at this region. We assumed that any of the following three factors can lead to a higher probability of a motif being bound by a TF and therefore taking the value 2: (i) more motifs of the same type present within a regulatory region; (ii) multiple TFs that can bind to the same motif, such as TAL1 and LYL1 both binding to Eboxes; (iii) higher expression levels of the TFs. The probabilities were thus calculated based on these three sources of information (see below for an example). We next defined that every regulatory region was a continuous variable on the close interval [0, 1], and its value was determined by the accumulated effects of all motifs present with the regulatory region. Finally, the expression levels of the nine TFs were also defined as continuous variables ranging from 0 to 0.8, and their expression levels were determined by the accumulated activities of the relevant regulatory regions.

Considering that variables in the top tier of the DBN are binary whereas those in the middle and bottom tiers are continuous, we found conditional linear Gaussian distribution (55) to be an appropriate generic representation of the intra-slice conditional probability distributions. Specifically, the regression coefficient of a regulatory region on a motif family was estimated by normalizing the logarithmic deviation of luciferase activity, where deviation refers to the

change of luciferase activity between the wild-type and the mutated (one motif family at a time) regulatory region (see below for a demonstration). Using the logarithmic deviation allowed us to account for the differences in effect sizes of various motif families by rescaling the differences to a comparable range. Similarly, for each of the nine genes, the regression coefficient of its expression level on a relevant regulatory region was estimated by normalizing the logarithmic deviation of luciferase activity, where deviation refers to the change of luciferase activity compared to the empty vector controls. All Matlab source codes are available at https://github.com/Huange and also http://burrn-sim.stemcells.cam.ac.uk/.

Detailed explanation of the modelling of each tier of the DBN:

a) Estimating the discrete probability distribution of a motif variable

The probability of a motif family at a given regulatory region taking value 1 or 2 (i.e., being unbound or bound) was calculated based on: (i) the number of such motifs in that regulatory region; (ii) the expression levels of the relevant TFs.

For example, three Ebox motifs were found at *Erg*+65 (Fig. 3a). They can be bound by either TAL1 or LYL1. Thus, we assigned that P(Ebox@Erg+65=1) and P(Ebox@Erg+65=2) were determined by {3, TAL1, LYL1}. We assumed that (i) the expression level of a TF is proportional to the probability of that TF binding to a target motif; and (ii) the bindings of TFs to multiple motifs are independent events. Gene expression levels were defined within the closed interval [0, 1], which is identical to the possible range of probabilities. For ease of calculation, we took the expression level of a TF as its probability of binding to a motif. Accordingly, we have

$$\tilde{P}(\text{Ebox@Erg}+65=1)=(1-p)^3\times(1-q)^3 \tag{1}$$

$$\tilde{P}(\text{Ebox@Erg}+65=2)=\sum_{n=1}^{3}C(3,n)\times p^n\times(1-p)^{(3-n)}\times(1-q)^3$$

$$+\sum_{n=1}^{3} C(3,n)\times q^{n}\times(1-q)^{(3-n)}\times(1-p)^{3} \quad (2)$$

$$+\sum_{n=1}^{2}\sum_{m=1}^{3-n} C(3,n)\times p^{n}\times(1-p)^{(3-n)}\times C\big((3-n),m\big)\times q^{m}\times(1-q)^{(3-m)}$$

where p and q represent the expression levels of TAL1 and LYL1, respectively.

However, to remove the bias introduced by simply taking the expression level of a TF as its probability of binding to a motif, we further normalized the resulting probabilities as below:

$$\tilde{Z} = \tilde{P}(\text{Ebox@Erg+65}=1) + \tilde{P}(\text{Ebox@Erg+65}=2) \quad (3)$$

$$P(\text{Ebox@Erg+65}=1) = \tilde{P}(\text{Ebox@Erg+65}=1)/\tilde{Z} \quad (4)$$

$$P(\text{Ebox@Erg+65}=2) = \tilde{P}(\text{Ebox@Erg+65}=2)/\tilde{Z} \quad (5)$$

It should be mentioned that the number of the same motifs in a regulatory region was directly taken into account in the estimation of probabilities. One may raise the question of whether this number has such strong power. Specifically, should the exponents in equations (1) and (2) change linearly, or less than linearly, along with the increase in the number of Ebox motifs? To address this issue, we replaced all exponents with their square roots and rerun the whole set of simulations (data not shown). Results showed that using the square roots instead of the original numbers (i) caused a more evenly distributed expression of the nine TFs over the hypothetical interval [0, 1], (ii) captured the same trend in gene expression changes in some perturbations (e.g. the AML-ETO simulation), but (iii) led to decreased expression levels of certain TFs in other perturbations (e.g. PU.1 knockdown and GFI1B over-expression), which therefore disagrees with the experimental data. In order to capture a better agreement of computational and experimental results, we directly used the number of motifs to estimate the discrete probability distributions.

b) Estimating the activity of a regulatory region

570 The regression coefficient of a regulatory region on a motif family was estimated by
571 normalizing the logarithmic deviation of luciferase activity, e.g. comparing the change of
572 luciferase activity between the wild-type and mutated constructs. For example, when the
573 luciferase activity for the wild-type *Erg*+65 region was set to 100 %, the simultaneous
574 mutation of all Ebox or Gfi motifs at this region resulted in increased luciferase activity (181.2
575 % or 475.9 %, respectively) (Fig. 3b). In contrast, simultaneous mutation of all Ets or Gata
576 motifs at this region led to reduced luciferase activity (1.3 % or 14.5 %, respectively). Based on
577 this information, we estimated the regression coefficient of the Erg+65 region on a relevant
578 motif family in the following way:

579 $$\alpha_i = \log\left(\frac{100}{l_k}\right) \times \left(\sum_k \left|\log\left(\frac{100}{l_k}\right)\right|\right)^{-1} \quad (6)$$

580 where $k \in \{1,...,4\}$, $l_1 = 181.2$, $l_2 = 475.9$, $l_3 = 1.3$, $l_4 = 14.5$; accordingly, $\alpha_1 = -0.070$,
581 $\alpha_2 = -0.185$, $\alpha_3 = 0.515$, $\alpha_4 = 0.230$. We can then formulate a linear regression equation as
582 below:

583 $$\tilde{y} = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 \quad (7)$$

584 where $\tilde{y}$ denote the estimated luciferase activity of *Erg*+65, and x_1, x_2, x_3 and x_4 represent
585 the binding status of Ebox, Gfi, Ets and Gata motifs at *Erg*+65.

586 However, the minimum and maximum $\tilde{y}$ obtained by the above formula are 0.235 (when
587 $x_1 = x_2 = 2$ and $x_3 = x_4 = 1$) and 1.235 (when $x_1 = x_2 = 1$ and $x_3 = x_4 = 2$). To make the values
588 of $\tilde{y}$ fall in the desired closed interval [0, 1], an intercept of -0.235 has to be introduced into the
589 linear regression model. In addition, a disturbance term has been included in the model in order
590 to satisfy the generic assumption of conditional linear Gaussian distribution. Finally, the fully
591 defined linear regression model regarding *Erg*+65 is given as:

592 $$\tilde{y} = c + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \varepsilon \quad (8)$$

where $c = -0.235$, $\varepsilon \sim N(0, \sigma^2)$, and σ should be a very small value.

c) Estimating the expression level of a gene

For each gene studied, the regression coefficient of its expression level on a relevant regulatory region was estimated by normalizing the logarithmic deviation of luciferase activity, where deviation refers to the change of luciferase activity compared to an empty vector control.

For example, when setting the luciferase activity of the wild-type constructs to 100 %, the luciferase activity of the empty vector controls relative to *Erg*+65, *Erg*+75 and *Erg*+85 wild-types are 1.9 %, 1.0 % and 15.2 %, respectively (Fig. 3b, Fig. 3-figure supplement 1 and 2). Based on these data, we estimated the expression level of Erg on a relevant regulatory region in the following way:

$$\beta_i = \log\left(\frac{100}{l_k}\right) \times \left(\sum_k \left|\log\left(\frac{100}{l_k}\right)\right|\right)^{-1} \quad (9)$$

where $k \in \{1,2,3\}$, $l_1 = 1.9$, $l_2 = 1.0$, $l_3 = 15.2$; accordingly, $\beta_1 = 0.379$, $\beta_2 = 0.441$, $\beta_3 = 0.180$. We can then formulate a linear regression equation as below:

$$\tilde{z} = \beta_1 \tilde{y}_1 + \beta_2 \tilde{y}_2 + \beta_3 \tilde{y}_3 \quad (10)$$

where $\tilde{z}$ denote the estimated expression level of Erg; and $\tilde{y}_1$, $\tilde{y}_2$ and $\tilde{y}_3$ represent the estimated activities of *Erg*+65, *Erg*+75 and *Erg*+85. Again, a disturbance term has been introduced to the model in order to meet the generic assumption of conditional linear Gaussian distribution. Thus, the fully defined linear regression model regarding *Erg* is given as:

$$\tilde{z} = \beta_1 \tilde{y}_1 + \beta_2 \tilde{y}_2 + \beta_3 \tilde{y}_3 + \varepsilon \quad (11)$$

where $\varepsilon \sim N(0, \sigma^2)$ and σ should be a very small value.

Statistics

Significance for the results of the luciferase reporter assays was calculated by combining the p-values of each experiment (generated by using the t-test function in Excel) using the Fisher's method, followed by the calculation of Stouffer's z trend if necessary. Significance tests for changes in TF expression levels caused by TF perturbations (both computational and experimental) were evaluated by Wilcoxon rank-sum tests.

Acknowledgments

We thank the CIMR Flow Cytometry Core facility, especially Dr Chiara Cossetti, for their expertise with cell sorting, Dr Marina Evangelou for her help with statistical analysis of the luciferase assay data, Lucas Greder for advice on cell transfection and stimulating discussions and Cyagen Biosciences and the MRC MHU Transgenic Core for generating transgenic embryos. Thanks are also extended to past and present members of the Göttgens and de Bruijn lab for practical assistance in the analysis of transient transgenic embryos, and to Yoram Groner and Ditsa Levanon for insightful discussions. We are grateful to Barbara L. Kee (University of Chicago, USA) for providing the MigRI-Gfi1b vector and Peter Laslo (University of Leeds, UK) for the shPU.1 construct.

Funding

Research in the authors' laboratories was supported by Bloodwise, The Wellcome Trust, Cancer Research UK, the Biotechnology and Biological Sciences Research Council, the National Institute of Health Research, the Medical Research Council, the MRC Molecular Haematology Unit (Oxford) core award, a Weizmann-UK "Making Connections" grant

(Oxford) and core support grants by the Wellcome Trust to the Cambridge Institute for Medical Research (100140) and Wellcome Trust–MRC Cambridge Stem Cell Institute (097922).

Competing Interests

The authors declare that no competing interests exist.

References

1. Davidson EH. Emerging properties of animal gene regulatory networks. Nature. 2010;468(7326):911-20.
2. Davidson EH. Network design principles from the sea urchin embryo. Current Opinion in Genetics & Development. 2009;19(6):535-40.
3. Petricka JJ, Benfey PN. Reconstructing regulatory network transitions. Trends Cell Biol. 2011;21(8):442-51.
4. Pimanda JE, Gottgens B. Gene regulatory networks governing haematopoietic stem cell development and identity. Int J Dev Biol. 2010;54(6-7):1201-11.
5. Gottgens B. Regulatory network control of blood stem cells. Blood. 2015;125(17):2614-20.
6. Schutte J, Moignard V, Gottgens B. Establishing the stem cell state: insights from regulatory network analysis of blood stem cell development. Wiley interdisciplinary reviews Systems biology and medicine. 2012;4(3):285-95.
7. Bonzanni N, Garg A, Feenstra KA, Schutte J, Kinston S, Miranda-Saavedra D, et al. Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. Bioinformatics (Oxford, England). 2013;29(13):i80-8.
8. Krumsiek J, Marr C, Schroeder T, Theis FJ. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. PloS one. 2011;6(8):e22649.
9. Narula J, Williams CJ, Tiwari A, Marks-Bluth J, Pimanda JE, Igoshin OA. Mathematical model of a gene regulatory network reconciles effects of genetic perturbations on hematopoietic stem cell emergence. Developmental Biology. 2013;379(2):258-69.
10. Narula J, Smith AM, Gottgens B, Igoshin OA. Modeling reveals bistability and low-pass filtering in the network module determining blood stem cell fate. PLoS computational biology. 2010;6(5):e1000771.
11. Gazit R, Garrison BS, Rao TN, Shay T, Costello J, Ericson J, et al. Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. Stem cell reports. 2013;1(3):266-80.
12. Jojic V, Shay T, Sylvia K, Zuk O, Sun X, Kang J, et al. Identification of transcriptional regulators in the mouse immune system. Nature immunology. 2013;14(6):633-43.
13. Dunn SJ, Martello G, Yordanov B, Emmott S, Smith AG. Defining an essential transcription factor program for naive pluripotency. Science (New York, NY). 2014;344(6188):1156-60.
14. Zhou Q, Chipperfield H, Melton DA, Wong WH. A gene regulatory network in mouse embryonic stem cells. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(42):16438-43.

15. Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P, et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell stem cell. 2010;7(4):532-44.
16. Wilson NK, Miranda-Saavedra D, Kinston S, Bonadies N, Foster SD, Calero-Nieto F, et al. The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. Blood. 2009;113(22):5456-65.
17. Beck D, Thoms JA, Perera D, Schutte J, Unnikrishnan A, Knezevic K, et al. Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. Blood. 2013;122(14):e12-22.
18. Pimanda JE, Ottersbach K, Knezevic K, Kinston S, Chan WY, Wilson NK, et al. Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(45):17692-7.
19. Wozniak RJ, Boyer ME, Grass JA, Lee Y, Bresnick EH. Context-dependent GATA factor function: combinatorial requirements for transcriptional control in hematopoietic and endothelial cells. The Journal of biological chemistry. 2007;282(19):14665-74.
20. Moignard V, Macaulay IC, Swiers G, Buettner F, Schutte J, Calero-Nieto FJ, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nature cell biology. 2013;15(4):363-72.
21. Chan WY, Follows GA, Lacaud G, Pimanda JE, Landry JR, Kinston S, et al. The paralogous hematopoietic regulators Lyl1 and Scl are coregulated by Ets and GATA factors, but Lyl1 cannot rescue the early Scl-/- phenotype. Blood. 2007;109(5):1908-16.
22. Wilkinson AC, Kawata VK, Schutte J, Gao X, Antoniou S, Baumann C, et al. Single-cell analyses of regulatory network perturbations using enhancer-targeting TALEs suggest novel roles for PU.1 during haematopoietic specification. Development. 2014;141(20):4018-30.
23. Nottingham WT, Jarratt A, Burgess M, Speck CL, Cheng JF, Prabhakar S, et al. Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. Blood. 2007;110(13):4188-97.
24. Gottgens B, Broccardo C, Sanchez MJ, Deveaux S, Murphy G, Gothert JR, et al. The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. Molecular and cellular biology. 2004;24(5):1870-83.
25. Gottgens B, Nastos A, Kinston S, Piltz S, Delabesse EC, Stanley M, et al. Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. The EMBO journal. 2002;21(12):3039-50.
26. Gottgens B, Ferreira R, Sanchez MJ, Ishibashi S, Li J, Spensberger D, et al. cis-Regulatory remodeling of the SCL locus during vertebrate evolution. Molecular and cellular biology. 2010;30(24):5741-51.
27. Calero-Nieto FJ, Ng FS, Wilson NK, Hannah R, Moignard V, Leal-Cervantes AI, et al. Key regulators control distinct transcriptional programmes in blood progenitor and mast cells. The EMBO journal. 2014;33(11):1212-26.
28. Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nature reviews Genetics. 2012;13(7):469-83.
29. Bee T, Ashley EL, Bickley SR, Jarratt A, Li PS, Sloane-Stanley J, et al. The mouse Runx1 +23 hematopoietic stem cell enhancer confers hematopoietic specificity to both Runx1 promoters. Blood. 2009;113(21):5121-4.
30. Sinclair AM, Gottgens B, Barton LM, Stanley ML, Pardanaud L, Klaine M, et al. Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. Dev Biol. 1999;209(1):128-42.
31. Sanchez M, Gottgens B, Sinclair AM, Stanley M, Begley CG, Hunter S, et al. An SCL 3' enhancer targets developing endothelium together with embryonic and adult haematopoietic progenitors. Development. 1999;126(17):3891-904.

32. Lelieveld SH, Schutte J, Dijkstra MJ, Bawono P, Kinston SJ, Gottgens B, et al. ConBind: motif-aware cross-species alignment for the identification of functional transcription factor binding sites. Nucleic acids research. 2015.
33. Mikkola HKA, Klintman J, Yang H, Hock H, Schlaeger TM, Fujiwara Y, et al. Haematopoietic stem cells retain long-term repopulating activity and multipotency in the absence of stem-cell leukaemia SCL/tal-1 gene. Nature. 2003;421(6922):547-51.
34. Hall MA, Curtis DJ, Metcalf D, Elefanty AG, Sourris K, Robb L, et al. The critical regulator of embryonic hematopoiesis, SCL, is vital in the adult for megakaryopoiesis, erythropoiesis, and lineage choice in CFU-S12. Proceedings of the National Academy of Sciences of the United States of America. 2003;100(3):992-7.
35. Capron C, Lecluse Y, Kaushik AL, Foudi A, Lacout C, Sekkai D, et al. The SCL relative LYL-1 is required for fetal and adult hematopoietic stem cell function and B-cell differentiation. Blood. 2006;107(12):4678-86.
36. Souroullas GP, Salmon JM, Sablitzky F, Curtis DJ, Goodell MA. Adult hematopoietic stem and progenitor cells require either Lyl1 or Scl for survival. Cell stem cell. 2009;4(2):180-6.
37. Licht JD. AML1 and the AML1-ETO fusion protein in the pathogenesis of t(8;21) AML. Oncogene. 2001;20(40):5660-79.
38. Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, et al. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Developmental cell. 2011;20(5):597-609.
39. Shivdasani RA, Mayer EL, Orkin SH. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. Nature. 1995;373(6513):432-4.
40. Taniuchi I, Sunshine MJ, Festenstein R, Littman DR. Evidence for distinct CD4 silencer functions at different stages of thymocyte differentiation. Molecular cell. 2002;10(5):1083-96.
41. Taniuchi I, Osato M, Egawa T, Sunshine MJ, Bae SC, Komori T, et al. Differential requirements for Runx proteins in CD4 repression and epigenetic silencing during T lymphocyte development. Cell. 2002;111(5):621-33.
42. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006;126(4):663-76.
43. Graf T, Enver T. Forcing cells to change lineages. Nature. 2009;462(7273):587-94.
44. Batta K, Florkowska M, Kouskoff V, Lacaud G. Direct reprogramming of murine fibroblasts to hematopoietic progenitor cells. Cell reports. 2014;9(5):1871-84.
45. Riddell J, Gazit R, Garrison BS, Guo G, Saadatpour A, Mandal PK, et al. Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. Cell. 2014;157(3):549-64.
46. Peter IS, Davidson EH. A gene regulatory network controlling the embryonic specification of endoderm. Nature. 2011;474(7353):635-9.
47. Pinto do OP, Kolterud A, Carlsson L. Expression of the LIM-homeobox gene LH2 generates immortalized steel factor-dependent multipotent hematopoietic precursors. The EMBO journal. 1998;17(19):5744-56.
48. Dexter TM, Allen TD, Scott D, Teich NM. Isolation and characterisation of a bipotential haematopoietic cell line. Nature. 1979;277(5696):471-4.
49. Busch K, Klapproth K, Barile M, Flossdorf M, Holland-Letz T, Schlenner SM, et al. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. Nature. 2015;518(7540):542-6.
50. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.
51. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
52. Sanchez-Castillo M, Ruau D, Wilkinson AC, Ng FS, Hannah R, Diamanti E, et al. CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. Nucleic acids research. 2015;43(Database issue):D1117-23.

53. Xu W, Kee BL. Growth factor independent 1B (Gfi1b) is an E2A target gene that modulates Gata3 in T-cell lymphomas. Blood. 2007;109(10):4406-14.
54. Wang W, Yang J, Liu H, Lu D, Chen X, Zenonos Z, et al. Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(45):18283-8.
55. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques.2009.
56. Longabaugh WJ, Davidson EH, Bolouri H. Computational representation of developmental genetic regulatory networks. Dev Biol. 2005;283(1):1-16.

Figure legends

Figure 1: Identification of haematopoietic active *cis*-regulatory regions. (a) UCSC screenshot of the *Erg* gene locus for ChIP-Sequencing data for nine haematopoietic TFs (ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 (15)) and for H3K27ac (27) in HPC7 cells. Highlighted are all regions of the *Erg* gene locus that are acetylated at H3K27 and are bound by three or more TFs. Numbers indicate the distance (in kb) from the ATG start codon. **(b)** Summary of the identification of candidate *cis*-regulatory regions for all nine TFs and subsequent analysis in transgenic mouse assays. The inspection of the nine gene loci and the application of the selection criteria (≥ 3 TFs bound and H3K27ac) identified a total of 49 candidate *cis*-regulatory regions. The heatmap shows the binding pattern of the nine TFs to all candidate regulatory elements in HPC7 cells: green = bound, grey = unbound. Haematopoietic activity in E11.5 transgenic mice is indicated by the font color: black = active, red = not active. Grey indicates genomic repeat regions that were not tested in transgenic mice. Detailed experimental data corresponding to the summary heatmap can be found in Fig. 1-figure supplement 1-8. **(c)** Haematopoietic activity of the five candidate *Erg cis*-regulatory regions was determined in E11.5 transgenic mouse assays. Shown are X-Gal-stained whole-mount embryos and paraffin sections of the dorsal aorta (DA, ventral side on the left/top) and foetal liver (FL), sites of definitive haematopoiesis. Colour coding as in B.

Figure 1 – figure supplement 1: Identification of haematopoietic active *cis*-regulatory elements for *Fli1*. (a) The candidate *cis*-regulatory elements were identified by ChIP-Seq analysis of the TFs ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 as well as H3K27 acetylation in the haematopoietic stem/progenitor cell line HPC7. Highlighted in pink are the candidate *cis*-regulatory regions which are bound by at least three of the nine TFs and showed H3K27 acetylation. The numbering represents the direction and distance in kilobases from the start codon ATG (pro = promoter). **(b)** Candidate regions were assayed for haematopoietic enhancer activity in mouse transient transgenic embryos. X-Gal stained whole-mount E11.5 embryos and paraffin sections of the dorsal aorta (DA; longitudinal section, ventral side on the left/top) and foetal liver (FL) are shown for the candidate *cis*-regulatory regions. Transgenic mouse data are not shown for previously published regions, but relevant publications are listed.

Figure 1 – figure supplement 2: Identification of haematopoietic active *cis*-regulatory elements for *Gata2*. (a) The candidate *cis*-regulatory elements were identified by ChIP-Seq analysis of the TFs ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 as well as H3K27 acetylation in the haematopoietic stem/progenitor cell line HPC7. Highlighted in pink are the candidate *cis*-regulatory regions which are bound by at least three of the nine TFs and showed H3K27 acetylation. The numbering represents the direction and distance in kilobases from the start codon ATG (pro = promoter). **(b)** Candidate regions were assayed for haematopoietic enhancer activity in mouse transient transgenic embryos. X-Gal stained whole-mount E11.5 embryos and paraffin sections of the dorsal aorta (DA; longitudinal section, ventral side on the left/top) and foetal liver (FL) are shown for the candidate *cis*-regulatory

regions. Transgenic mouse data are not shown for previously published regions, but relevant publications are listed.

Figure 1 – figure supplement 3: Identification of haematopoietic active *cis*-regulatory elements for *Gfi1b*. (a) The candidate *cis*-regulatory elements were identified by ChIP-Seq analysis of the TFs ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 as well as H3K27 acetylation in the haematopoietic stem/progenitor cell line HPC7. Highlighted in pink are the candidate *cis*-regulatory regions which are bound by at least three of the nine TFs and showed H3K27 acetylation. The numbering represents the direction and distance in kilobases from the start codon ATG (pro = promoter). **(b)** All candidate regions were previously published regions. Relevant publications are listed.

Figure 1 – figure supplement 4: Identification of haematopoietic active *cis*-regulatory elements for *Lyl1*. (a) The candidate *cis*-regulatory elements were identified by ChIP-Seq analysis of the TFs ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 as well as H3K27 acetylation in the haematopoietic stem/progenitor cell line HPC7. Highlighted in pink are the candidate *cis*-regulatory regions which are bound by at least three of the nine TFs and showed H3K27 acetylation. The numbering represents the direction and distance in kilobases from the start codon ATG (pro = promoter). **(b)** Candidate regions were assayed for haematopoietic enhancer activity in mouse transient transgenic embryos. X-Gal stained whole-mount E11.5 embryos and paraffin sections of the dorsal aorta (DA; longitudinal section, ventral side on the left/top) and foetal liver (FL) are shown for the candidate *cis*-regulatory regions. Transgenic mouse data are not shown for previously published regions, but relevant publications are listed.

Figure 1 – figure supplement 5: Identification of haematopoietic active *cis*-regulatory elements for *Meis1*. (a) The candidate *cis*-regulatory elements were identified by ChIP-Seq analysis of the TFs ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 as well as H3K27 acetylation in the haematopoietic stem/progenitor cell line HPC7. Highlighted in pink are the candidate *cis*-regulatory regions which are bound by at least three of the nine TFs and showed H3K27 acetylation. The numbering represents the direction and distance in kilobases from the start codon ATG (pro = promoter). **(b)** Candidate regions were assayed for haematopoietic enhancer activity in mouse transient transgenic embryos. X-Gal stained whole-mount E11.5 embryos and paraffin sections of the dorsal aorta (DA; longitudinal section, ventral side on the left/top) and foetal liver (FL) are shown for the candidate *cis*-regulatory regions.

Figure 1 – figure supplement 6: Identification of haematopoietic active *cis*-regulatory elements for *Runx1*. (a) The candidate *cis*-regulatory elements were identified by ChIP-Seq analysis of the TFs ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 as well as H3K27 acetylation in the haematopoietic stem/progenitor cell line HPC7. Highlighted in pink are the candidate *cis*-regulatory regions which are bound by at least three of the nine TFs and showed H3K27 acetylation. The numbering represents the direction and distance in kilobases from the start codon ATG (pro = promoter). **(b)** E10 embryos and cryosections of the DA (transverse; ventral down) and FL are shown. For the *Runx1*+204 region, a larger 12 kb fragment (chr16:92,620,915-92,631,936, mm9) was used for transient transgenesis, but similar results were obtained with the +204 fragment alone (data not shown). The +24 element was tested in conjunction with the +23 and did not change its tissue specificity (Bee et al., 2010).

Preliminary data show that the +24 on its own does not mediate robust tissue specific expression of reporter genes. Transgenic mouse data are not shown for previously published regions, but relevant publications are listed.

Figure 1 – figure supplement 7: Identification of haematopoietic active *cis*-regulatory elements for *Spi1*. (a) The candidate *cis*-regulatory elements were identified by ChIP-Seq analysis of the TFs ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 as well as H3K27 acetylation in the haematopoietic stem/progenitor cell line HPC7. Highlighted in pink are the candidate *cis*-regulatory regions which are bound by at least three of the nine TFs and showed H3K27 acetylation. The numbering represents the direction and distance in kilobases from the start codon ATG (pro = promoter). **(b)** All candidate regions were previously published regions. Relevant publications are listed.

Figure 1 – figure supplement 8: Identification of haematopoietic active *cis*-regulatory elements for *Tal1*. (a) The candidate *cis*-regulatory elements were identified by ChIP-Seq analysis of the TFs ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1 as well as H3K27 acetylation in the haematopoietic stem/progenitor cell line HPC7. Highlighted in pink are the candidate *cis*-regulatory regions which are bound by at least three of the nine TFs and showed H3K27 acetylation. The numbering is based on the distance (in kb) to promoter 1a. **(b)** All candidate regions were previously published regions. Relevant publications are listed.

Figure 1 – figure supplement 9: Number of PCR and LacZ positive transgenic embryos (E10.5-11.5) for each regulatory region.

Figure 2: Comparison of TF binding pattern at haematopoietic active *cis*-regulatory regions in two haematopoietic progenitor cell lines, HPC7 and 416b. (a) UCSC screenshot of the *Erg* gene locus for ChIP-Sequencing data for nine haematopoietic TFs (ERG, FLI1, GATA2, GFI1B, LYL1, MEIS1, PU.1, RUNX1 and TAL1) and for H3K27ac in 416b cells. Highlighted are those haematopoietic active *Erg cis*-regulatory regions that were identified based on acetylation of H3K27 and TF binding in HPC7 cells followed by transgenic mouse assays. Numbers indicate the distance (in kb) from the ATG start codon. **(b)** Hierarchical clustering of the binding profiles for HPC7, 416b and other published datasets. The heatmap shows the pairwise correlation coefficient of peak coverage data between pairs of samples in the row and column. The order of the samples is identical in columns and rows. Details about samples listed can be found in Fig. 2-figure supplement 9. **(c)** Pair-wise analysis of binding of the nine TFs to haematopoietic active *cis*-regulatory regions of the nine TFs in HPC7 versus 416b cells. Green = bound in both cells types, blue = only bound in 416b cells, orange = only bound in HPC7 cells, grey = not bound in either cell type.

Figure 2 – figure supplement 1: UCSC screenshot for the *Fli1* gene locus demonstrating binding patterns for nine key haematopoietic TFs and H3K27ac in 416b cells. Highlighted in pink are *cis*-regulatory regions that were identified based on the selection criteria (≥ 3 TFs bound and H3K27ac) in HPC7 cells and were shown to possess haematopoietic activity. The numbering represents the distance (in kb) from the start codon ATG.

Figure 2 – figure supplement 2: UCSC screenshot for the *Gata2* gene locus demonstrating binding patterns for nine key haematopoietic TFs and H3K27ac in 416b cells. Highlighted

928 in pink are *cis*-regulatory regions that were identified based on the selection criteria (≥ 3 TFs
929 bound and H3K27ac) in HPC7 cells and were shown to possess haematopoietic activity. The
930 numbering represents the distance (in kb) from the start codon ATG.

931

932 **Figure 2 – figure supplement 3: UCSC screenshot for the *Gfi1b* gene locus demonstrating**
933 **binding patterns for nine key haematopoietic TFs and H3K27ac in 416b cells.** Highlighted
934 in pink are *cis*-regulatory regions that were identified based on the selection criteria (≥ 3 TFs
935 bound and H3K27ac) in HPC7 cells and were shown to possess haematopoietic activity. The
936 numbering represents the distance (in kb) from the start codon ATG.

937

938 **Figure 2 – figure supplement 4: UCSC screenshot for the *Lyl1* gene locus demonstrating**
939 **binding patterns for nine key haematopoietic TFs and H3K27ac in 416b cells.** Highlighted
940 in pink is the promoter ("pro") that was identified based on the selection criteria (≥ 3 TFs
941 bound and H3K27ac) in HPC7 cells and was shown to possess haematopoietic activity. The
942 promoter is labelled with "pro".

943

944 **Figure 2 – figure supplement 5: UCSC screenshot for the *Meis1* gene locus demonstrating**
945 **binding patterns for nine key haematopoietic TFs and H3K27ac in 416b cells.** Highlighted
946 in pink is the *cis*-regulatory region that was identified based on the selection criteria (≥ 3 TFs
947 bound and H3K27ac) in HPC7 cells and was shown to possess haematopoietic activity. The
948 numbering represents the distance (in kb) from the start codon ATG.

949

Figure 2 – figure supplement 6: UCSC screenshot for the *Runx1* gene locus demonstrating binding patterns for nine key haematopoietic TFs and H3K27ac in 416b cells. Highlighted in pink are *cis*-regulatory regions that were identified based on the selection criteria (≥ 3 TFs bound and H3K27ac) in HPC7 cells and were subsequently shown to possess haematopoietic activity. The numbering represents the distance (in kb) from the start codon ATG.

Figure 2 – figure supplement 7: UCSC screenshot for the *Spi1* gene locus demonstrating binding patterns for nine key haematopoietic TFs and H3K27ac in 416b cells. Highlighted in pink is the *cis*-regulatory region that was identified based on the selection criteria (≥ 3 TFs bound and H3K27ac) in HPC7 cells and was shown to possess haematopoietic activity. The numbering represents the distance (in kb) from the start codon ATG.

Figure 2 – figure supplement 8: UCSC screenshot for the *Tal1* gene locus demonstrating binding patterns for nine key haematopoietic TFs and H3K27ac in 416b cells. Highlighted in pink are *cis*-regulatory regions that were identified based on the selection criteria (≥ 3 TFs bound and H3K27ac) in HPC7 cells and were shown to possess haematopoietic activity. The numbering the distance (in kb) from promoter 1a.

Figure 2 – figure supplement 9: List of ChIP-Seq samples included in the heatmap in Figure 2b.

Figure 3: TFBS mutagenesis reveals enhancer-dependent effects of TF binding on gene expression. (a) Multiple species alignment of mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1) sequences for the *Erg*+65 region. Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata, yellow = Gfi, red = Meis. The nucleotides changed to mutate the TFBSs are indicated below the alignment. All binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** Luciferase assay for the *Erg*+65 wild-type and mutant enhancer in stably transfected 416b cells. Each bar represents the averages of at least three independent experiments with three to four replicates within each experiment. Results are shown relative to the wild-type enhancer activity, which is set to 100%. Error bars represent the standard error of the mean (SEM). Stars indicate significance: ** = p-value < 0.01, *** = p-value < 0.001. P-values were calculated using t-tests, followed by the Fisher's method. **(c)** Summary of luciferase assay results for all 19 high-confidence haematopoietic active regulatory regions. Relative luciferase activity is illustrated in shades of blue (down-regulation) and red (up-regulation). Crossed-out grey boxes indicate that there is no motif for the TF and/or the TF does not bind to the region. Detailed results and corresponding alignments with highlighted TFBSs and their mutations can be found in Figure 3-figure supplements 1-18.

Figure 3 – figure supplement 1: Multiple species alignment and luciferase assay results for *Erg*+75. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are

highlighted in: blue = Ebox, purple = Ets, yellow = Gfi. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. Where TF binding was observed in ChIP-Seq experiments in 416b cells, but the TFBS was not conserved, the motifs present in the mouse sequence only were mutated. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: ** = p-value < 0.01, *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 2: Multiple species alignment and luciferase assay results for *Erg*+85. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata, yellow = Gfi. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 3: Multiple species alignment and luciferase assay results for *Fli1*+12. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 4: Multiple species alignment and luciferase assay results for *Gata2*-93. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata, red = Meis, turquoise = Runt. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars

indicate significance: ** = p-value < 0.01, *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 5: Multiple species alignment and luciferase assay results for *Gata2*+3. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 6: Multiple species alignment and luciferase assay results for *Gfi1b*+16. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata, yellow = Gfi, red = Meis, turquoise = Runt. The nucleotides that were changed to mutate the TFBSs are indicated

below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. Where TF binding was observed in ChIP-Seq experiments in 416b cells, but the TFBS was not conserved, the motifs present in the mouse sequence only were mutated. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: ** = p-value < 0.01, *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 7: Multiple species alignment and luciferase assay results for *Gfi1b*+17. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata, yellow = Gfi, red = Meis. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: ** = p-value < 0.01, *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 8: Multiple species alignment and luciferase assay results for *Lyl1* promoter. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2) and opossum (monDom5). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between three of four species. Transcription factor binding sites (TFBS) are highlighted in: purple = Ets, green = Gata. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ets motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 9: Multiple species alignment and luciferase assay results for *Meis1*+48. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: purple = Ets, green = Gata, yellow = Gfi, red = Meis. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ets motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: *** =

p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 10: Multiple species alignment and luciferase assay results for *Spi1*-14. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, turquoise = Runt. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: ** = p-value < 0.01, *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 11: Multiple species alignment and luciferase assay results for *Runx1*-59. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19) and dog (canFam2). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between two of three species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata, red = Meis. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox

motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 12: Multiple species alignment and luciferase assay results for *Runx1*+3. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata, yellow = Gfi, red = Meis, turquoise = Runt. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ets motifs) were mutated simultaneously. Where TF binding was observed in ChIP-Seq experiments in 416b cells, but the TFBS was not conserved, the motifs present in the mouse sequence only were mutated. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: * = p-value < 0.05, ** = p-value < 0.01, *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 13: Multiple species alignment and luciferase assay results for *Runx1*+23. (a) Multiple species alignment (MSA) with the following species: mouse

(mm9), human (hg19), dog (canFam2) and opossum (monDom5). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between three to four species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata, red = Meis, turquoise = Runt. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: * = p-value < 0.05, *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 14: Multiple species alignment and luciferase assay results for *Runx1*+110. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ets motifs) were mutated simultaneously. Where TF binding was observed in ChIP-Seq experiments in 416b cells, but the TFBS was not conserved, the motifs present in the mouse sequence only were mutated. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: ** = p-value < 0.01, *** = p-value < 0.001. P-

values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 15: Multiple species alignment and luciferase assay results for *Runx1*+204. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2), opossum (monDom5) and platypus (ornAna1). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between four of five species. Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, yellow = Gfi, turquoise = Runt. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ets motifs) were mutated simultaneously. Where TF binding was observed in ChIP-Seq experiments in 416b cells, but the TFBS was not conserved, the motifs present in the mouse sequence only were mutated. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 16: Multiple species alignment and luciferase assay results for *Tal1*-4. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19) and dog (canFam2). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between two of three species. Transcription factor binding sites (TFBS) are highlighted in: purple = Ets. The nucleotides that

were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of the Ets motif family were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: ** = p-value < 0.01. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 17: Multiple species alignment and luciferase assay results for *Tal1*+19. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19), dog (canFam2) and opossum (monDom5). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between three of four species. Transcription factor binding sites (TFBS) are highlighted in: purple = Ets. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of the Ets motif family were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: *** = p-value < 0.001. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 18: Multiple species alignment and luciferase assay results for *Tal1*+40. (a) Multiple species alignment (MSA) with the following species: mouse (mm9), human (hg19) and dog (canFam2). Nucleotides highlighted in black are conserved between all species analysed, nucleotides highlighted in grey are conserved between two of three species.

Transcription factor binding sites (TFBS) are highlighted in: blue = Ebox, purple = Ets, green = Gata. The nucleotides that were changed to mutate the TFBSs are indicated below the MSA. All conserved binding sites of one motif family (e.g. all Ebox motifs) were mutated simultaneously. **(b)** For the luciferase reporter assays in stably transfected 416b cells the averages of at least three independent experiments with three to four replicates within each experiment are shown. Error bars represent the standard error of the mean (SEM). Stars indicate significance: * = p-value < 0.05. P-values were generated using t-tests, followed by the Fisher's method and if necessary Stouffer's z trend.

Figure 3 – figure supplement 19: List of TF binding sites and the TFs that bind to them.

Figure 3 – figure supplement 20: List of co-ordinates and primer sequences for the regulatory regions analysed in this study.

Figure 4: A three-tier DBN incorporating transcriptional regulatory information can recapitulate the HSPC expression state. (a) Representation of the complete network diagram generated using the Biotapestry software (56). **(b)** Schematic diagram describing the DBN which contains three tiers: I. TF binding motifs within regulatory regions, II. *cis*-regulatory regions influencing the expression levels of the various TFs, and III. genes encoding the TFs. The output of tier III, namely the expression levels of the TF, feed back into the TF binding at the various motifs of tier I. The model therefore is comprised of successive time slices (t). **(c)** Simulation of a single cell over time. The expression levels of all 9 TFs are the same at the beginning (0.5). The simulation rapidly stabilizes with characteristic TF expression levels. (**d**)

Simulation of a cell population by running the model 1000 times. The scale of the x-axis is linear. Each simulation was run as described in (c).

Figure 4 - figure supplement 1: Simulation of a single cell over time with different expression levels at the beginning. The simulation rapidly stabilizes with characteristic TF expression levels irrespective of the starting conditions. (**a**) The expression levels of all 9 TFs are 0.2 at the start of the simulation. (**b**) The expression levels of all 9 TFs are 0.8 at the start of the simulation. (**c**) The expression levels for FLI1, RUNX1 and TAL1 are set to be 0.5 at the beginning, with all other TFs not being expressed (value of 0).

Figure 5: The DBN recapitulates the consequences of TAL1 and LYL1 single and double perturbations as seen *in vivo* and *in vitro*. Computational prediction of gene expression patterns for the nine TFs of interest after perturbation of TAL1 (**a**), LYL1 (**b**) or both (**c**). Deletion of TAL1 or LYL1 on their own has no major consequences on the expression levels of the other eight TFs of the gene regulatory network, but simultaneous deletion of both TAL1 and LYL1 caused changes in expression of several genes, mainly a decrease in *Gata2* and *Runx1*. This major disruption of the core GRN for blood stem/progenitor cells is therefore consistent with TAL1/LYL1 double knockout HSCs showing a much more severe phenotype than the respective single knock-outs. One thousand simulations were run for each perturbation to determine the TFs expression levels in a "cell population" by selecting expression levels at random time points after reaching its initial steady state. Expression levels of 0 resemble no expression, whereas expression levels of 1 stand for highest expression level that is possible in this system. The scale of the x-axis is linear. (**d**) Gene expression levels measured in single 416b cells transfected with siRNA constructs against Tal1 or a control. The density plots of

gene expression levels after perturbation of TAL1 indicate the relative number of cells (y-axes) at each expression level (x-axes). The scale of the x-axis is linear. The values indicate the results of the Wilcoxon rank-sum test: alterations to the expression profiles are indicated by the p-value (statistical significance: $p < 0.001$ for computational data and $p < 0.05$ for experimental data); substantial shifts in median expression level are indicated by the shift of median (SOM) (SOM >0.1 for computational data and >1 for experimental data). For details, see Fig. 5 – figure supplement 1; for full expression data, see figure 5 – source data.

Fig. 5 - figure supplement 1: Significance tests for the computational and experimental data after TF perturbations. To determine statistical significance the Wilcoxon rank-sum test was used. Alterations to the expression profiles are indicated by the p-value; with statistically significance defined as follows: $p < 0.001$ for computational data and $p < 0.05$ for experimental data. Significance of a substantial shift in median expression levels are as follows: shift of median >0.1 for computational data and >1 for experimental data (because of different scales). If the number for the shift of median is negative, the median of the perturbation data is smaller than that of the wild-type control; if the number is positive, the median of the perturbation is larger than that of the control. For simplicity, all significant changes are highlighted in red (p-value) and blue (shift of median).

Figure 5 - figure supplement 2: Histogram plots showing the gene expression distributions of all nine genes of the network for the following perturbations: (**a**) LYL1 down-regulation; (**b**) TAL1/SCL down-regulation; (**c**) LYL1 and TAL1/SCL down-regulation; (**d**) PU.1 down-regulation; (**e**) GFI1B up-regulation; and (**f**) AML-ETO9a simulation.

Figure 5 - source data: Raw and normalised data for the single cell gene expression experiments presented in this study: 1) TAL1 down-regulation (related to Fig. 5 d), 2) PU.1 down-regulation (related to Fig. 6 a), 3) GFI1B up-regulation (related to Fig. 6b) and 4) AML-ETO9a perturbation (related to Fig. 6 c)

Figure 6: The DBN captures the transcriptional consequences of network perturbations. Left panel: Computational prediction of gene expression after perturbation of specific TFs. 1000 simulations were run for each perturbation to determine expression levels in a "cell population" (expression at 0 resembles no expression, whereas expression of 1 represents the highest possible expression level). The scale of the x-axis is linear. **Right panel:** Density plots of gene expression levels in single 416b cells after perturbation of specific TFs indicating the relative number of cells at each expression level. The scale of the x-axis is linear. The values indicate the results of the Wilcoxon rank-sum test: alterations to the expression profiles are indicated by the p-value (statistical significance: $p < 0.001$ for computational data and $p < 0.05$ for experimental data); substantial shifts in median expression level are indicated by the shift of median (SOM) (SOM >0.1 for computational data and >1 for experimental data). For details, see Fig. 5 – figure supplement 1. **(a)** PU.1 down-regulation: (Left) Computational prediction of gene expression after PU.1 knockdown (Spi1 was set to 0 after reaching its initial steady state). (Right) Gene expression levels measured in single 416b cells transduced with shRNA constructs against shluc (wild-type) or shPU.1 (PU.1 knockdown). **(b)** GFI1B over-expression: (Left) Computational prediction of gene expression after over-expression of GFI1B (Gfi1b was set to 1 after reaching its initial steady state). (Right) Gene expression levels in single 416b cells transduced with a Gfi1b-expressing vector compared to an empty vector control (wild-type). **(c)** Consequences of the AML-ETO9a oncogene: (Left) Computational prediction of gene expression patterns after introducing the dominant-negative effect of the AML-ETO9a

oncogene (Runx1 was fixed at the maximum value of 1 after reaching its initial steady state and in addition all Runt binding sites were set to have a repressive effect). (Right) Gene expression levels measured in single 416b cells transduced with an AML-ETO9a expressing vector fused to mCherry. mCherry positive cells were compared to mCherry negative cells (wild-type).

Figure 6 – figure supplement 1: Summary of all computational simulations for perturbations of one or two TFs. The results for a total of 162 simulations are shown. The data can be accessed using the embedded hyperlinks. The y-axes show the number of cells and the x-axes the relative expression level. Blue curves represent wild-type data and red curves represent perturbation data.

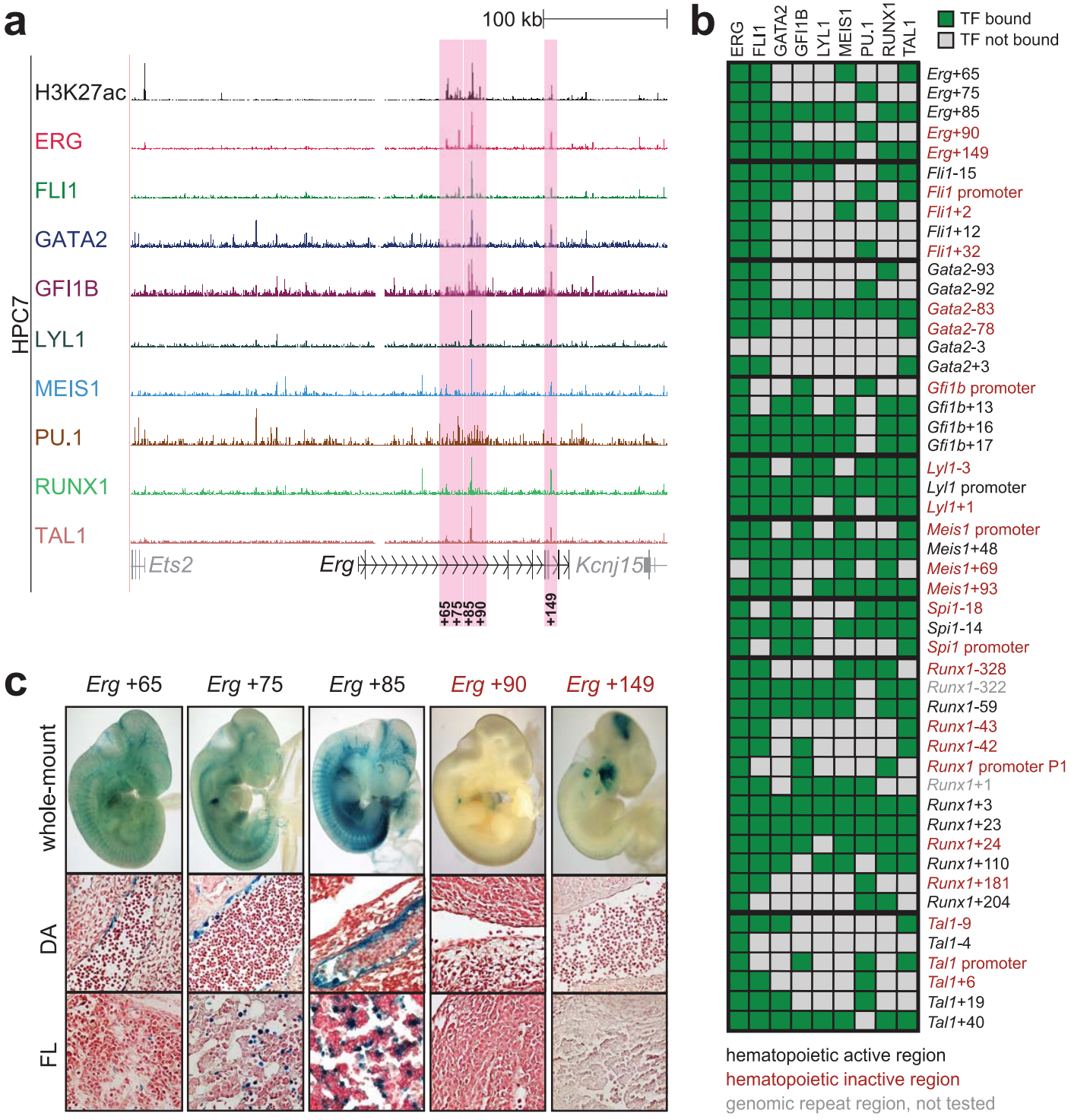
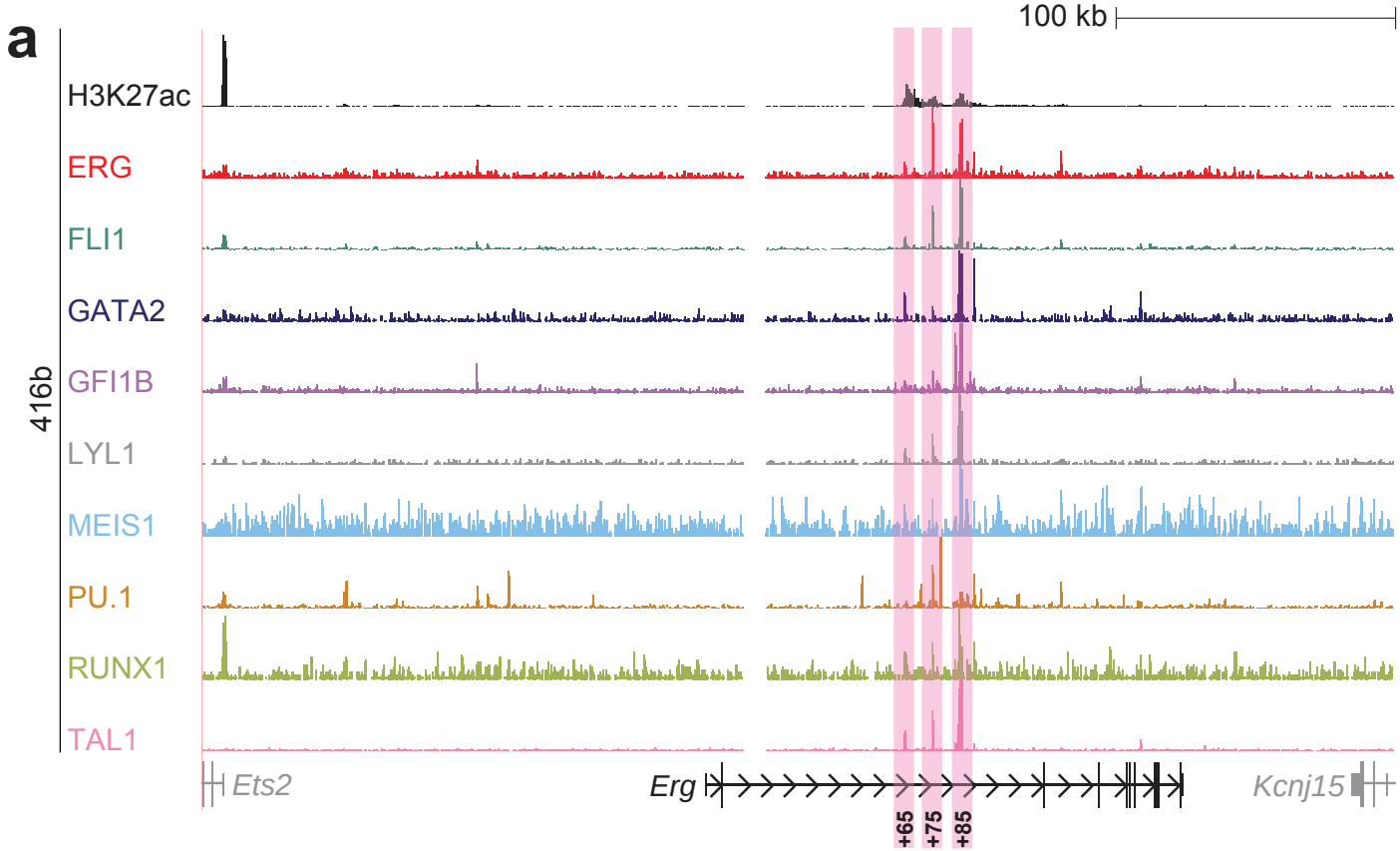
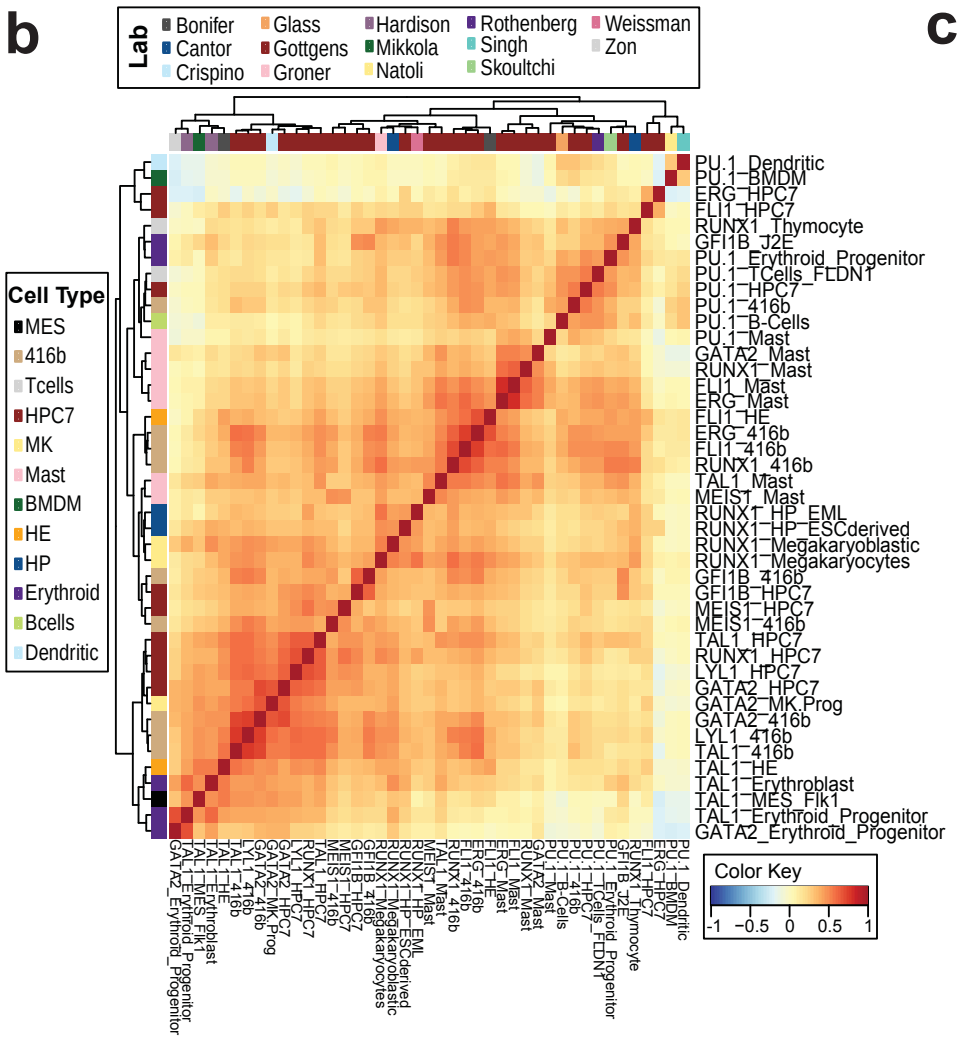

a
100 kb
HPC7
H3K27ac
ERG
FLI1
GATA2
GFI1B
LYL1
MEIS1
PU.1
RUNX1
TAL1
Ets2
Erg
Kcnj15
+65
+75
+85
+90
+149
b
ERG
FLI1
GATA2
GFI1B
LYL1
MEIS1
PU.1
RUNX1
TAL1
TF bound
TF not bound
Erg+65
Erg+75
Erg+85
Erg+90
Erg+149
Fli1-15
Fli1 promoter
Fli1+2
Fli1+12
Fli1+32
Gata2-93
Gata2-92
Gata2-83
Gata2-78
Gata2-3
Gata2+3
Gfi1b promoter
Gfi1b+13
Gfi1b+16
Gfi1b+17
Lyl1-3
Lyl1 promoter
Lyl1+1
Meis1 promoter
Meis1+48
Meis1+69
Meis1+93
Spi1-18
Spi1-14
Spi1 promoter
Runx1-328
Runx1-322
Runx1-59
Runx1-43
Runx1-42
Runx1 promoter P1
Runx1+1
Runx1+3
Runx1+23
Runx1+24
Runx1+110
Runx1+181
Runx1+204
Tal1-9
Tal1-4
Tal1 promoter
Tal1+6
Tal1+19
Tal1+40
hematopoietic active region
hematopoietic inactive region
genomic repeat region, not tested
c
Erg +65
Erg +75
Erg +85
Erg +90
Erg +149
whole-mount
DA
FL

Figure 1


a
100 kb
H3K27ac
ERG
FLI1
GATA2
GFI1B
LYL1
MEIS1
PU.1
RUNX1
TAL1
416b
Ets2
Erg
Kcnj15
+65
+75
+85


b
Lab
Bonifer
Cantor
Crispino
Glass
Gottgens
Groner
Hardison
Mikkola
Natoli
Rothenberg
Singh
Skoultchi
Weissman
Zon
Cell Type
MES
416b
Tcells
HPC7
MK
Mast
BMDM
HE
HP
Erythroid
Bcells
Dendritic
PU.1_Dendritic
PU.1_BMDM
ERG_HPC7
FLI1_HPC7
RUNX1_Thymocyte
GFI1B_J2E
PU.1_Erythroid_Progenitor
PU.1_TCells_FLDN1
PU.1_HPC7
PU.1_416b
PU.1_B-Cells
PU.1_Mast
GATA2_Mast
RUNX1_Mast
FLI1_Mast
ERG_Mast
FLI1_HE
ERG_416b
FLI1_416b
RUNX1_416b
TAL1_Mast
MEIS1_Mast
RUNX1_HP_EML
RUNX1_HP_ESCderived
RUNX1_Megakaryoblastic
RUNX1_Megakaryocytes
GFI1B_416b
GFI1B_HPC7
MEIS1_HPC7
MEIS1_416b
TAL1_HPC7
RUNX1_HPC7
LYL1_HPC7
GATA2_HPC7
GATA2_MK.Prog
GATA2_416b
LYL1_416b
TAL1_416b
TAL1_HE
TAL1_Erythroblast
TAL1_MES_Flk1
TAL1_Erythroid_Progenitor
GATA2_Erythroid_Progenitor
Color Key
-1
-0.5
0
0.5
1

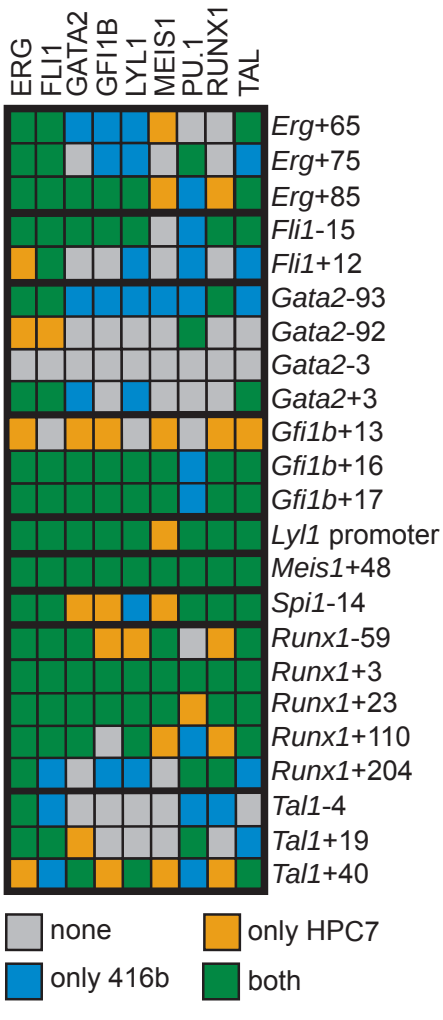

c
ERG
FLI1
GATA2
GFI1B
LYL1
MEIS1
PU.1
RUNX1
TAL
Erg+65
Erg+75
Erg+85
Fli1-15
Fli1+12
Gata2-93
Gata2-92
Gata2-3
Gata2+3
Gfi1b+13
Gfi1b+16
Gfi1b+17
Lyl1 promoter
Meis1+48
Spi1-14
Runx1-59
Runx1+3
Runx1+23
Runx1+110
Runx1+204
Tal1-4
Tal1+19
Tal1+40
none
only HPC7
only 416b
both

Figure 2

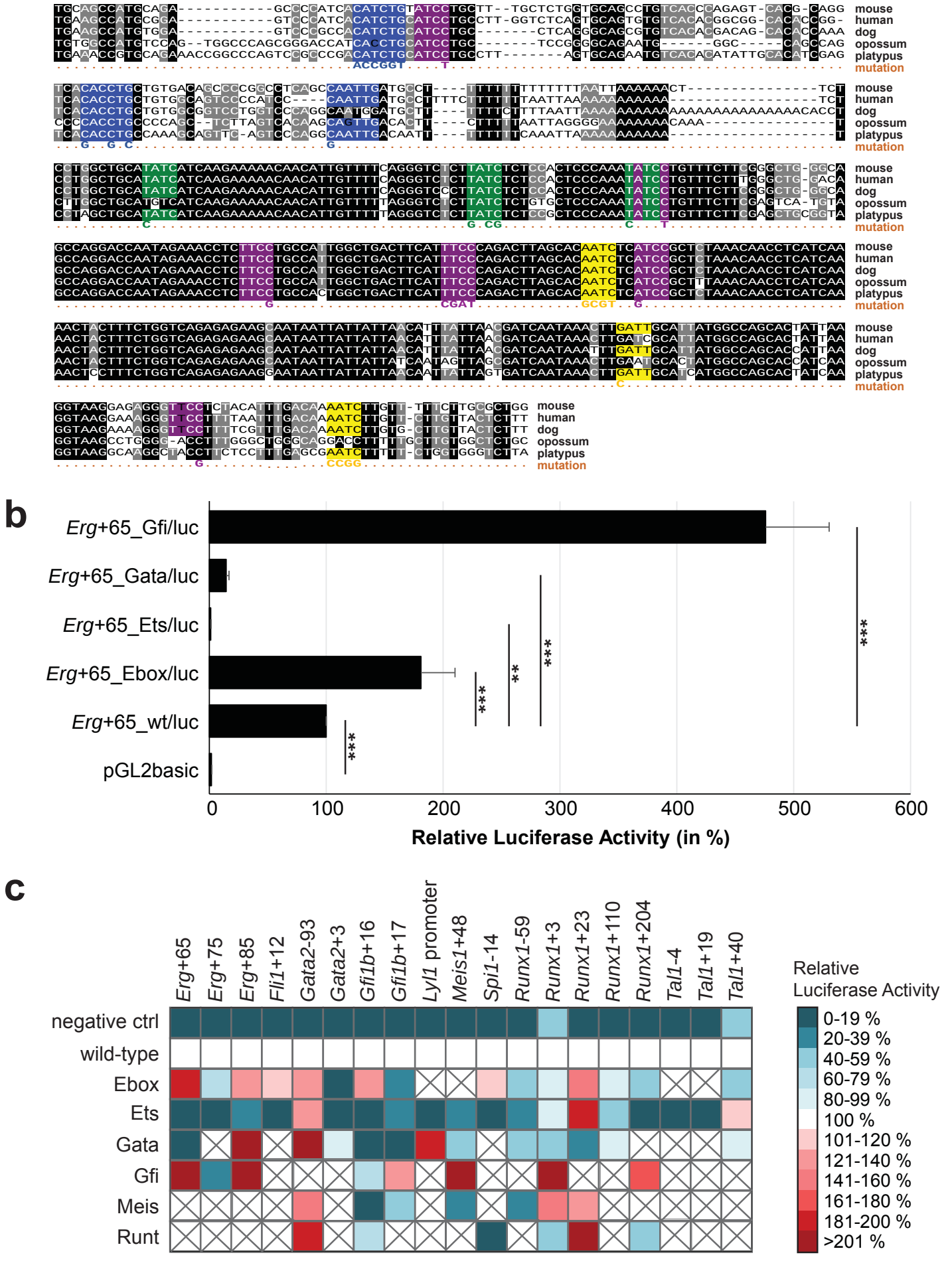

a
Erg+65
mouse
human
dog
opossum
platypus
mutation
b
Erg+65_Gfi/luc
Erg+65_Gata/luc
Erg+65_Ets/luc
Erg+65_Ebox/luc
Erg+65_wt/luc
pGL2basic
***
**
0
100
200
300
400
500
600
Relative Luciferase Activity (in %)
c
Erg+65
Erg+75
Erg+85
Fli1+12
Gata2-93
Gata2+3
Gfi1b+16
Gfi1b+17
Lyl1 promoter
Meis1+48
Spi1-14
Runx1-59
Runx1+3
Runx1+23
Runx1+110
Runx1+204
Tal1-4
Tal1+19
Tal1+40
negative ctrl
wild-type
Ebox
Ets
Gata
Gfi
Meis
Runt
Relative Luciferase Activity
0-19 %
20-39 %
40-59 %
60-79 %
80-99 %
100 %
101-120 %
121-140 %
141-160 %
161-180 %
181-200 %
>201 %

Figure 3

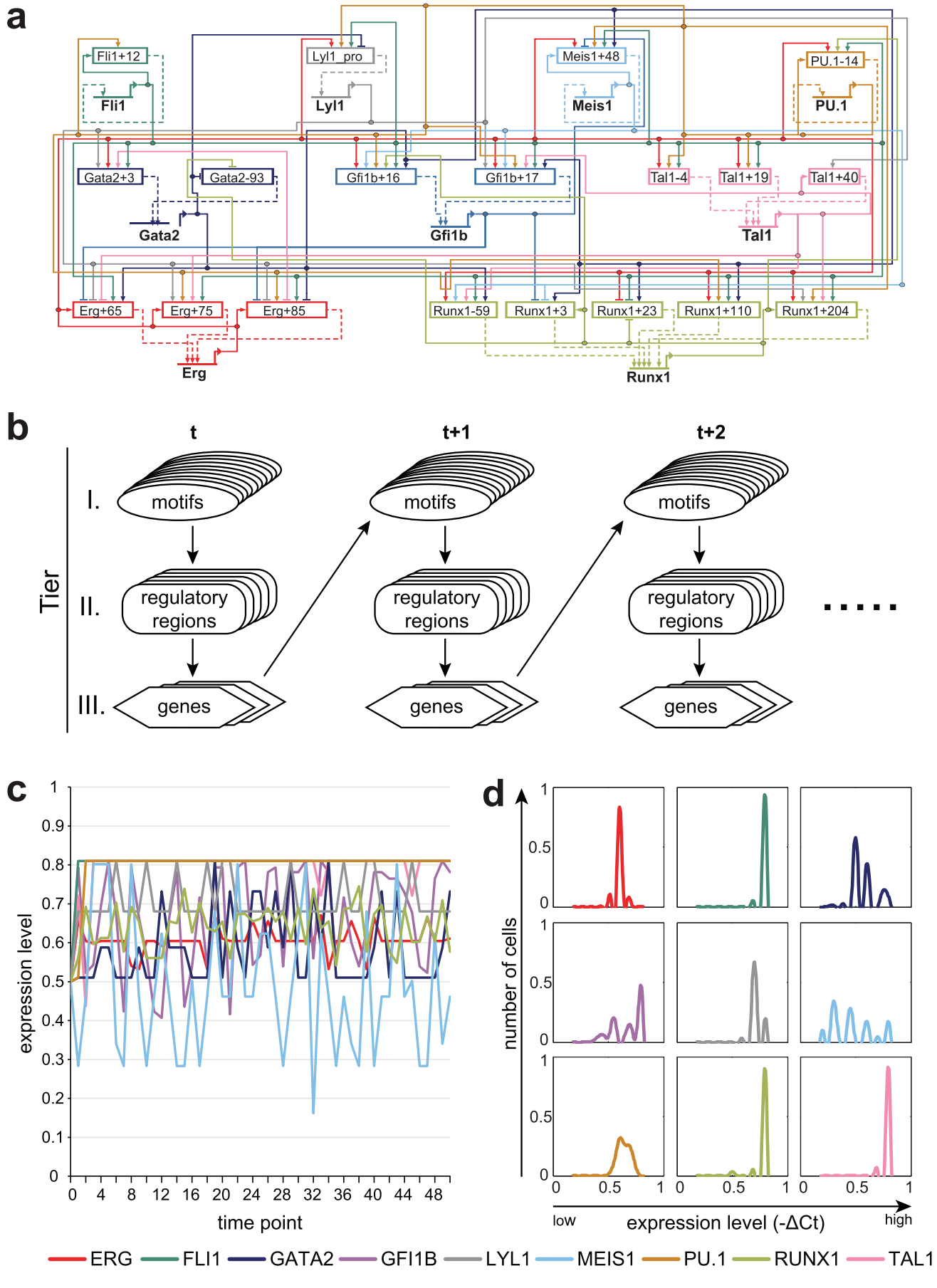

a
Fli1+12
Fli1
Lyl1_pro
Lyl1
Meis1+48
Meis1
PU.1-14
PU.1
Gata2+3
Gata2-93
Gata2
Gfi1b+16
Gfi1b+17
Gfi1b
Tal1-4
Tal1+19
Tal1+40
Tal1
Erg+65
Erg+75
Erg+85
Erg
Runx1-59
Runx1+3
Runx1+23
Runx1+110
Runx1+204
Runx1
b
t
t+1
t+2
Tier
I.
motifs
II.
regulatory regions
III.
genes
c
expression level
time point
d
number of cells
low
expression level (-ΔCt)
high
ERG
FLI1
GATA2
GFI1B
LYL1
MEIS1
PU.1
RUNX1
TAL1

Figure 4

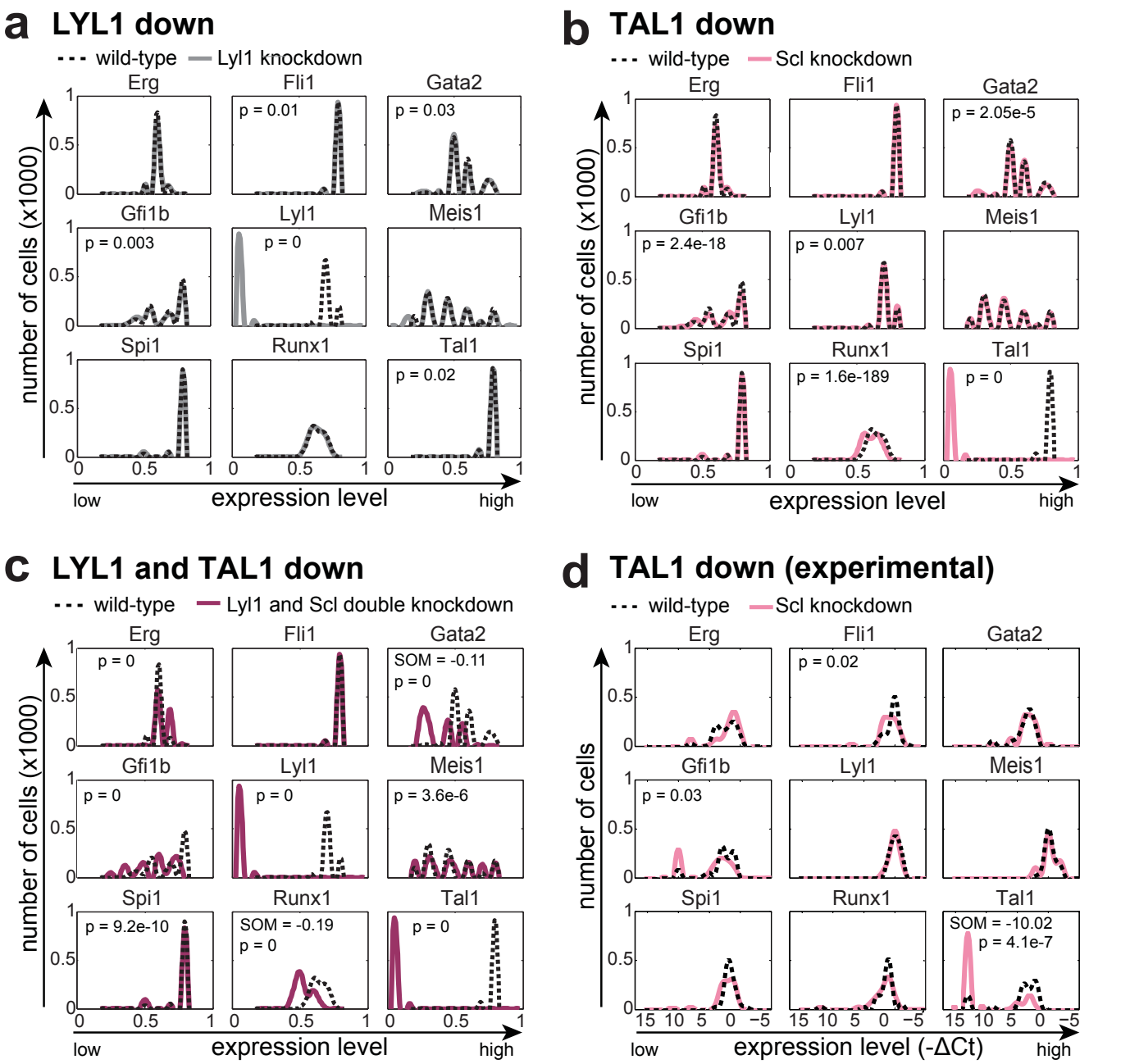

a LYL1 down
wild-type
Lyl1 knockdown
Erg
Fli1
Gata2
p = 0.01
p = 0.03
Gfi1b
Lyl1
Meis1
p = 0.003
p = 0
Spi1
Runx1
Tal1
p = 0.02
number of cells (x1000)
low
expression level
high
b TAL1 down
wild-type
Scl knockdown
Erg
Fli1
Gata2
p = 2.05e-5
Gfi1b
Lyl1
Meis1
p = 2.4e-18
p = 0.007
Spi1
Runx1
Tal1
p = 1.6e-189
p = 0
number of cells (x1000)
low
expression level
high
c LYL1 and TAL1 down
wild-type
Lyl1 and Scl double knockdown
Erg
Fli1
Gata2
p = 0
SOM = -0.11
p = 0
Gfi1b
Lyl1
Meis1
p = 0
p = 0
p = 3.6e-6
Spi1
Runx1
Tal1
p = 9.2e-10
SOM = -0.19
p = 0
p = 0
number of cells (x1000)
low
expression level
high
d TAL1 down (experimental)
wild-type
Scl knockdown
Erg
Fli1
Gata2
p = 0.02
Gfi1b
Lyl1
Meis1
p = 0.03
Spi1
Runx1
Tal1
SOM = -10.02
p = 4.1e-7
number of cells
low
expression level (-ΔCt)
high

Figure 5

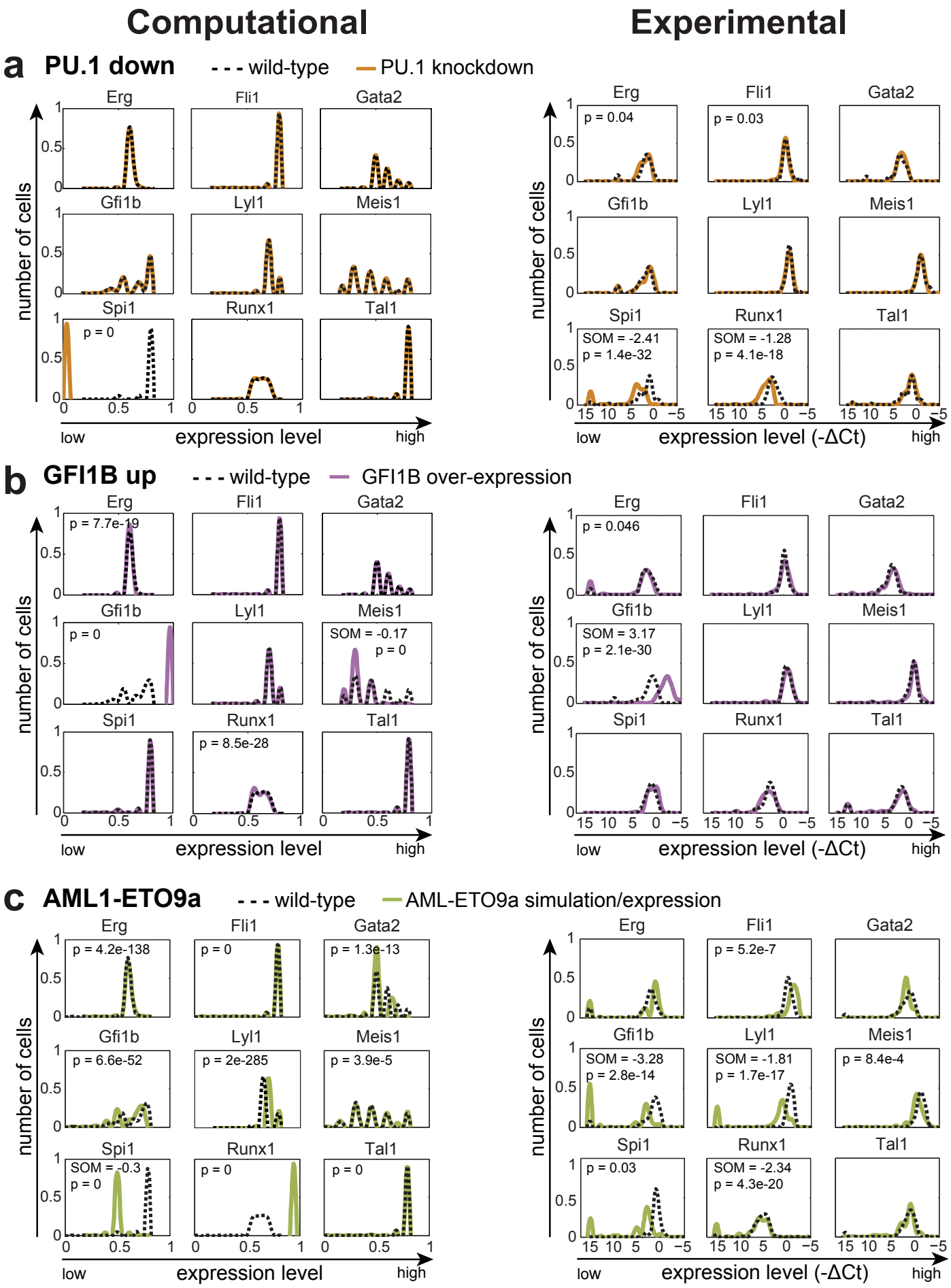

Computational
Experimental
a PU.1 down
- - - wild-type
PU.1 knockdown
b GFI1B up
- - - wild-type
GFI1B over-expression
c AML1-ETO9a
- - - wild-type
AML-ETO9a simulation/expression
Erg
Fli1
Gata2
Gfi1b
Lyl1
Meis1
Spi1
Runx1
Tal1
number of cells
low
expression level
high
expression level (-ΔCt)
p = 0
p = 0.04
p = 0.03
SOM = -2.41
p = 1.4e-32
SOM = -1.28
p = 4.1e-18
p = 7.7e-19
p = 0.046
SOM = -0.17
p = 0
SOM = 3.17
p = 2.1e-30
p = 8.5e-28
p = 4.2e-138
p = 1.3e-13
p = 5.2e-7
p = 6.6e-52
p = 2e-285
p = 3.9e-5
SOM = -3.28
p = 2.8e-14
SOM = -1.81
p = 1.7e-17
p = 8.4e-4
SOM = -0.3
p = 0.03
SOM = -2.34
p = 4.3e-20

Figure 6