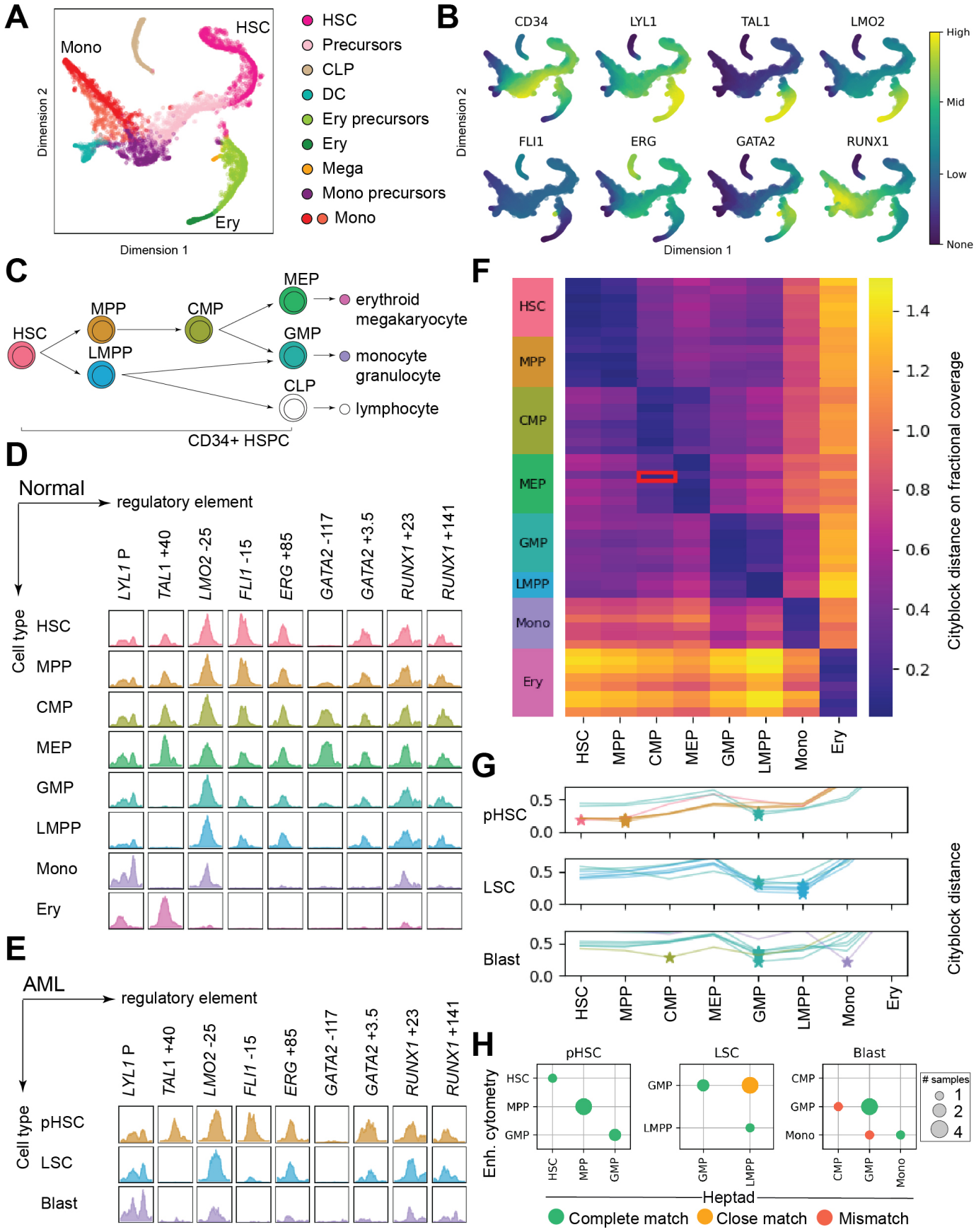
A
HSC
Mono
Ery
Dimension 1
Dimension 2
HSC
Precursors
CLP
DC
Ery precursors
Ery
Mega
Mono precursors
Mono
B
CD34
LYL1
TAL1
LMO2
FLI1
ERG
GATA2
RUNX1
High
Mid
Low
None
Dimension 1
Dimension 2
C
HSC
MPP
LMPP
CMP
MEP
GMP
CLP
erythroid megakaryocyte
monocyte granulocyte
lymphocyte
CD34+ HSPC
D
Normal
regulatory element
LYL1 P
TAL1 +40
LMO2 -25
FLI1 -15
ERG +85
GATA2 -117
GATA2 +3.5
RUNX1 +23
RUNX1 +141
Cell type
HSC
MPP
CMP
MEP
GMP
LMPP
Mono
Ery
E
AML
regulatory element
LYL1 P
TAL1 +40
LMO2 -25
FLI1 -15
ERG +85
GATA2 -117
GATA2 +3.5
RUNX1 +23
RUNX1 +141
Cell type
pHSC
LSC
Blast
F
HSC
MPP
CMP
MEP
GMP
LMPP
Mono
Ery
Cityblock distance on fractional coverage
1.4
1.2
1.0
0.8
0.6
0.4
0.2
G
pHSC
LSC
Blast
0.5
0.0
Cityblock distance
H
pHSC
LSC
Blast
Enh. cytometry
samples
1
2
4
Heptad
Complete match
Close match
Mismatch

Figure 1

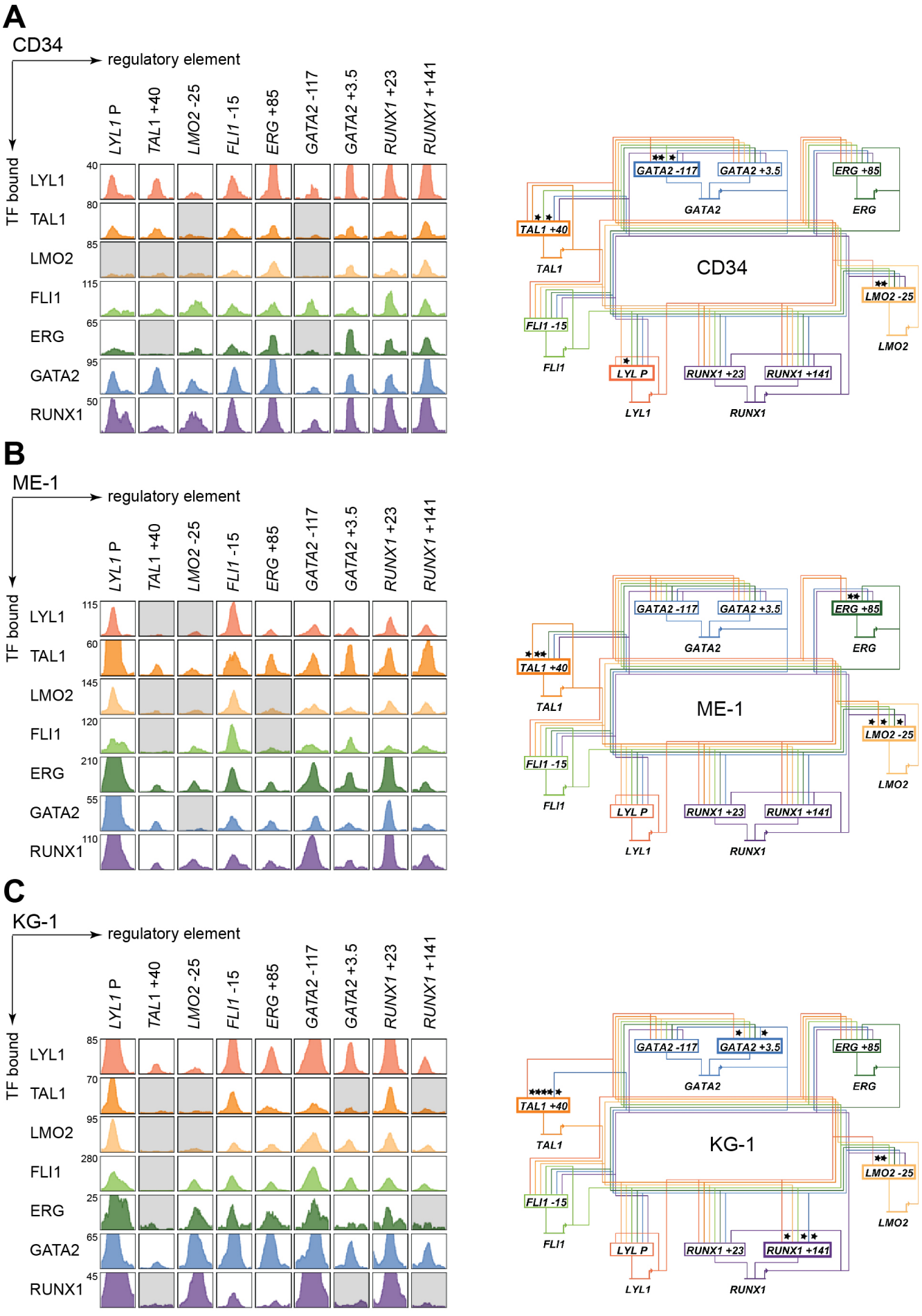
A
CD34
regulatory element
LYL1 P
TAL1 +40
LMO2 -25
FLI1 -15
ERG +85
GATA2 -117
GATA2 +3.5
RUNX1 +23
RUNX1 +141
TF bound
LYL1
TAL1
LMO2
FLI1
ERG
GATA2
RUNX1
40
80
85
115
65
95
50
GATA2 -117
GATA2 +3.5
GATA2
ERG +85
ERG
TAL1 +40
TAL1
CD34
LMO2 -25
LMO2
FLI1 -15
FLI1
LYL P
RUNX1 +23
RUNX1 +141
LYL1
RUNX1
B
ME-1
regulatory element
LYL1 P
TAL1 +40
LMO2 -25
FLI1 -15
ERG +85
GATA2 -117
GATA2 +3.5
RUNX1 +23
RUNX1 +141
TF bound
LYL1
TAL1
LMO2
FLI1
ERG
GATA2
RUNX1
115
60
145
120
210
55
110
GATA2 -117
GATA2 +3.5
GATA2
ERG +85
ERG
TAL1 +40
TAL1
ME-1
LMO2 -25
LMO2
FLI1 -15
FLI1
LYL P
RUNX1 +23
RUNX1 +141
LYL1
RUNX1
C
KG-1
regulatory element
LYL1 P
TAL1 +40
LMO2 -25
FLI1 -15
ERG +85
GATA2 -117
GATA2 +3.5
RUNX1 +23
RUNX1 +141
TF bound
LYL1
TAL1
LMO2
FLI1
ERG
GATA2
RUNX1
85
70
95
280
25
65
45
GATA2 -117
GATA2 +3.5
GATA2
ERG +85
ERG
TAL1 +40
TAL1
KG-1
LMO2 -25
LMO2
FLI1 -15
FLI1
LYL P
RUNX1 +23
RUNX1 +141
LYL1
RUNX1

Figure 2

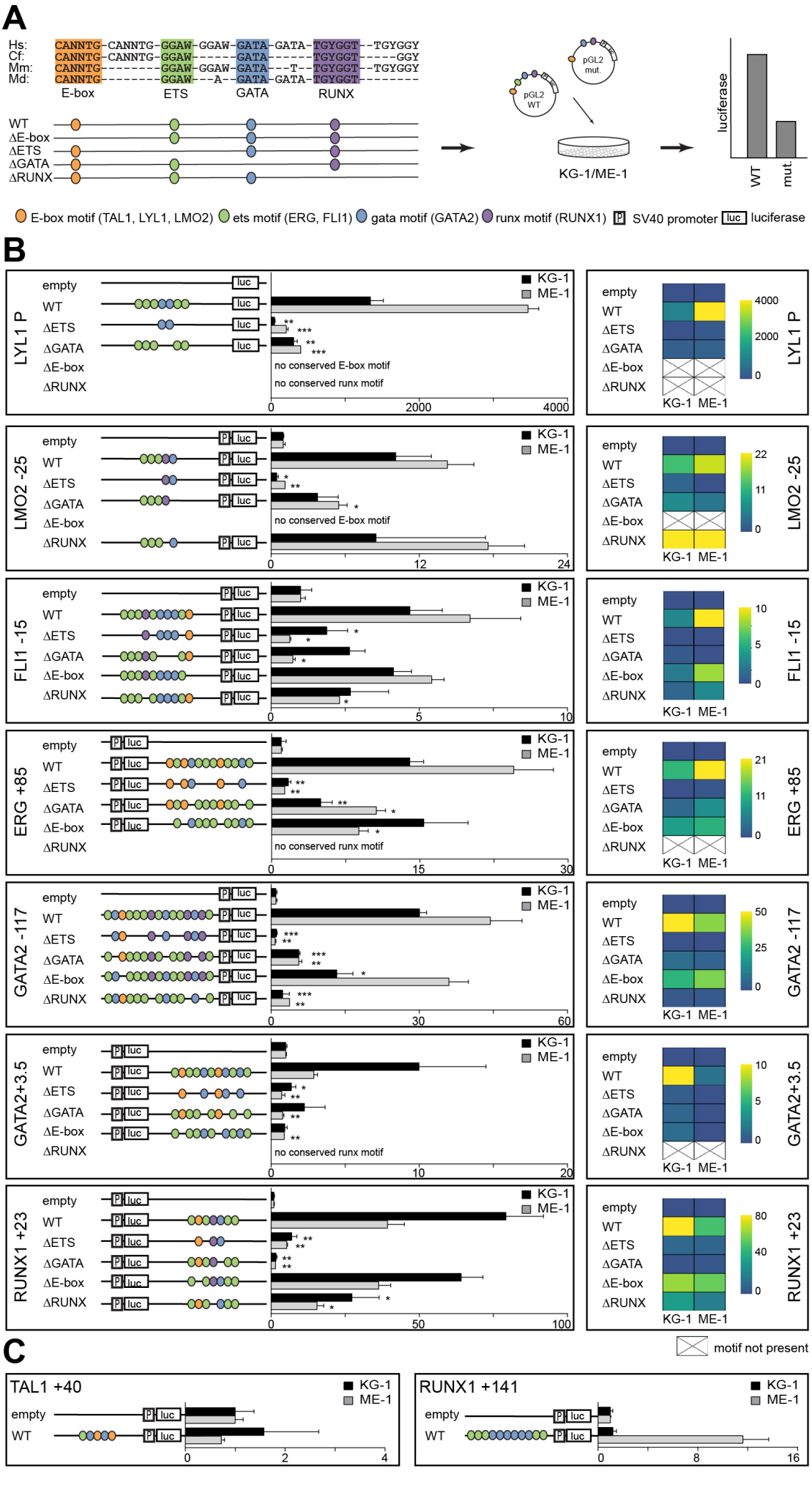
A
Hs: CANNTG-CANNTG-GGAW-GGAW-GATA-GATA-TGYGGT--TGYGGY
Cf: CANNTG-CANNTG-GGAW------GATA------TGYGGT-----GGY
Mm: CANNTG--------GGAW-GGAW-GATA---T--TGYGGT--TGYGGY
Md: CANNTG--------GGAW---A--GATA-GATA-TGYGGT--------
E-box
ETS
GATA
RUNX
WT
ΔE-box
ΔETS
ΔGATA
ΔRUNX
pGL2 WT
pGL2 mut.
KG-1/ME-1
luciferase
WT
mut.
E-box motif (TAL1, LYL1, LMO2)
ets motif (ERG, FLI1)
gata motif (GATA2)
runx motif (RUNX1)
P SV40 promoter
luc luciferase
B
LYL1 P
LMO2 -25
FLI1 -15
ERG +85
GATA2 -117
GATA2+3.5
RUNX1 +23
empty
WT
ΔETS
ΔGATA
ΔE-box
ΔRUNX
KG-1
ME-1
no conserved E-box motif
no conserved runx motif
motif not present
C
TAL1 +40
RUNX1 +141

Figure 3

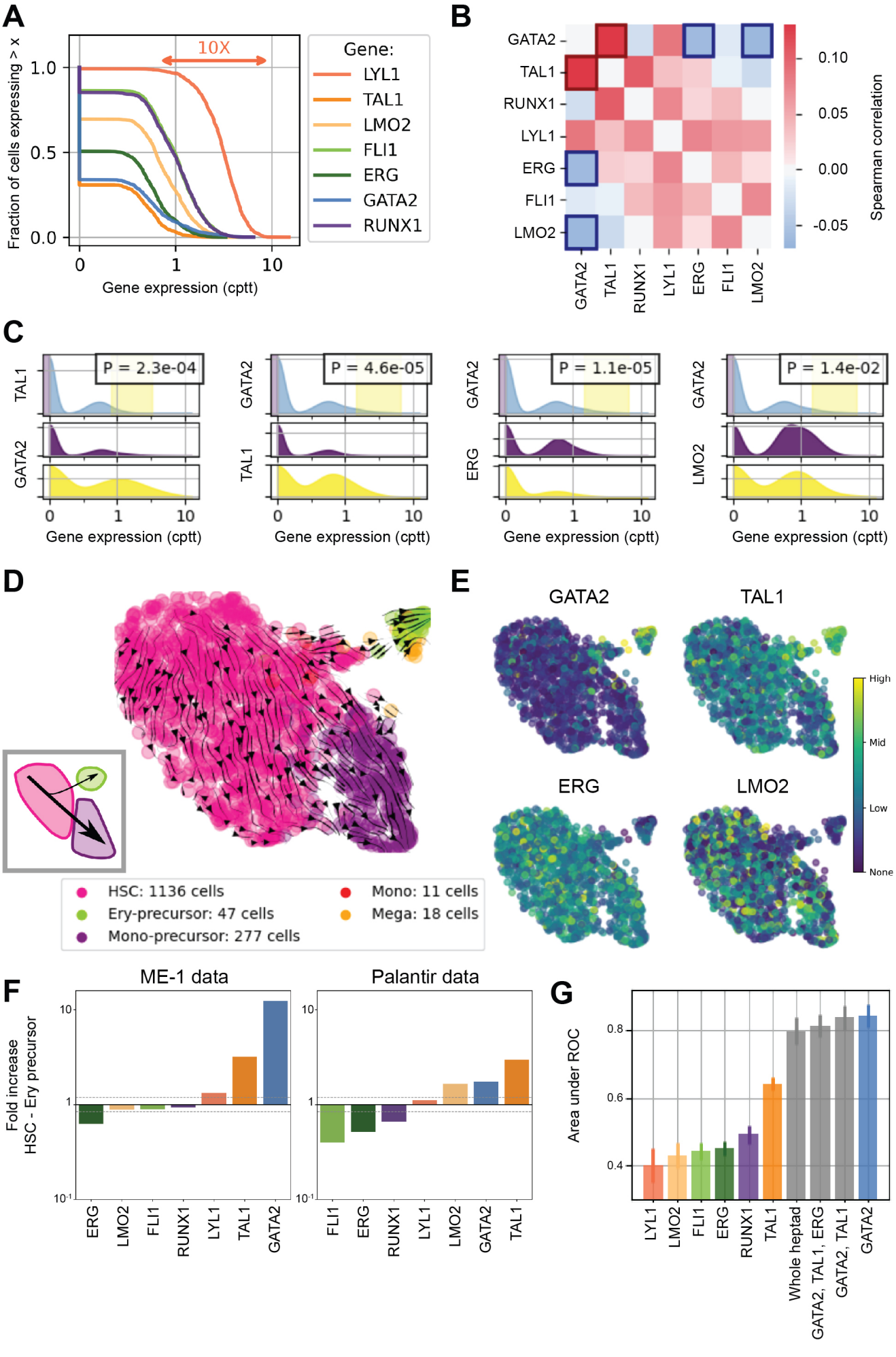
A
Fraction of cells expressing > x
1.0
0.5
0.0
10X
0
1
10
Gene expression (cptt)
Gene:
LYL1
TAL1
LMO2
FLI1
ERG
GATA2
RUNX1
B
GATA2
TAL1
RUNX1
LYL1
ERG
FLI1
LMO2
0.10
0.05
0.00
-0.05
Spearman correlation
C
TAL1
GATA2
P = 2.3e-04
P = 4.6e-05
P = 1.1e-05
P = 1.4e-02
ERG
LMO2
Gene expression (cptt)
D
HSC: 1136 cells
Ery-precursor: 47 cells
Mono-precursor: 277 cells
Mono: 11 cells
Mega: 18 cells
E
GATA2
TAL1
ERG
LMO2
High
Mid
Low
None
F
ME-1 data
Palantir data
Fold increase
HSC - Ery precursor
10
1
10⁻¹
ERG
LMO2
FLI1
RUNX1
LYL1
TAL1
GATA2
G
Area under ROC
0.8
0.6
0.4
LYL1
LMO2
FLI1
ERG
RUNX1
TAL1
Whole heptad
GATA2, TAL1, ERG
GATA2, TAL1
GATA2

Figure 4

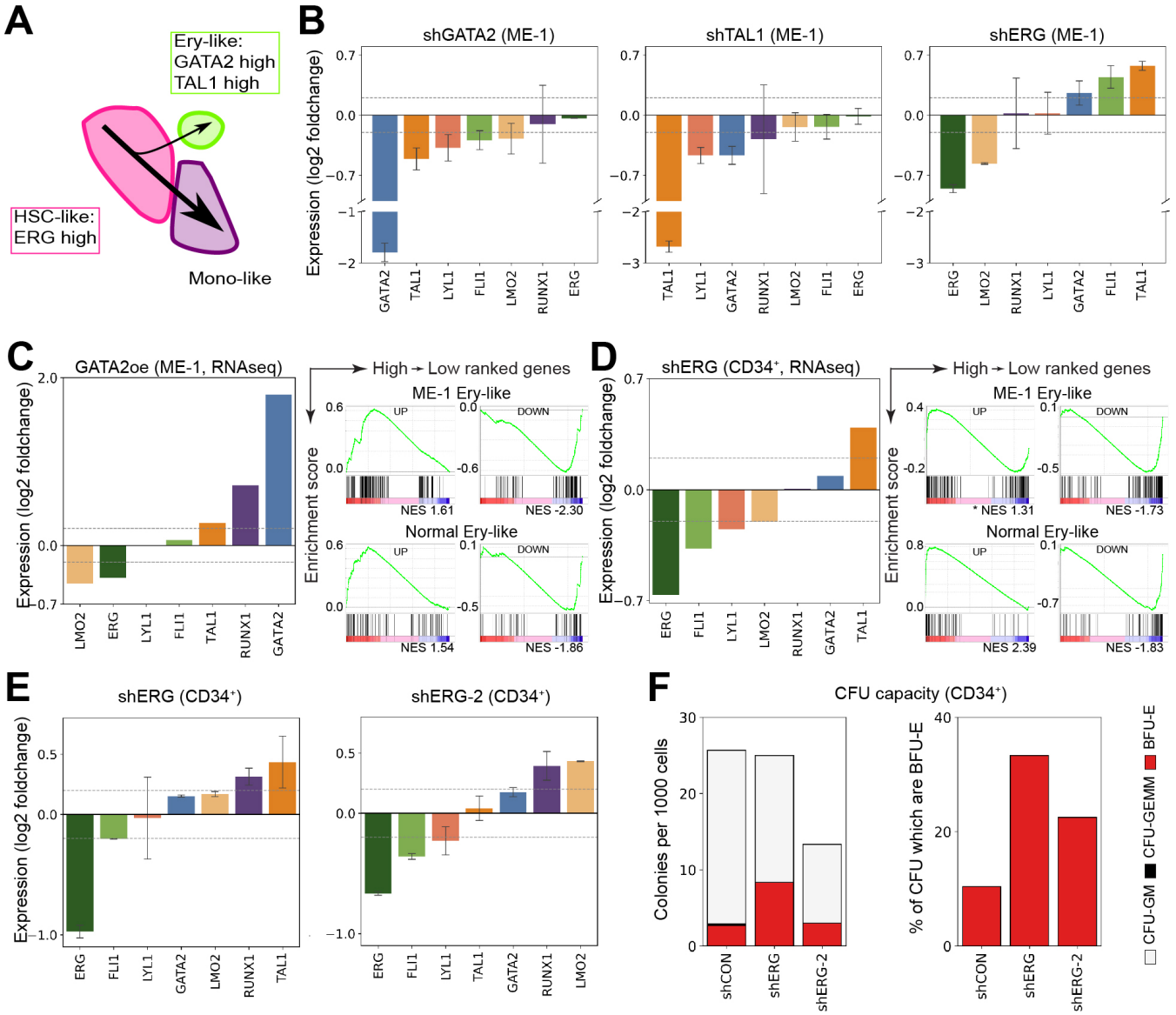
A
Ery-like:
GATA2 high
TAL1 high
HSC-like:
ERG high
Mono-like
B
shGATA2 (ME-1)
shTAL1 (ME-1)
shERG (ME-1)
Expression (log2 foldchange)
GATA2 TAL1 LYL1 FLI1 LMO2 RUNX1 ERG
TAL1 LYL1 GATA2 RUNX1 LMO2 FLI1 ERG
ERG LMO2 RUNX1 LYL1 GATA2 FLI1 TAL1
C
GATA2oe (ME-1, RNAseq)
Expression (log2 foldchange)
LMO2 ERG LYL1 FLI1 TAL1 RUNX1 GATA2
High → Low ranked genes
ME-1 Ery-like
Enrichment score
UP
DOWN
NES 1.61
NES -2.30
Normal Ery-like
NES 1.54
NES -1.86
D
shERG (CD34+, RNAseq)
Expression (log2 foldchange)
ERG FLI1 LYL1 LMO2 RUNX1 GATA2 TAL1
High → Low ranked genes
ME-1 Ery-like
Enrichment score
UP
DOWN
* NES 1.31
NES -1.73
Normal Ery-like
NES 2.39
NES -1.83
E
shERG (CD34+)
shERG-2 (CD34+)
Expression (log2 foldchange)
ERG FLI1 LYL1 GATA2 LMO2 RUNX1 TAL1
ERG FLI1 LYL1 TAL1 GATA2 RUNX1 LMO2
F
CFU capacity (CD34+)
Colonies per 1000 cells
% of CFU which are BFU-E
shCON shERG shERG-2
CFU-GM CFU-GEMM BFU-E

Figure 5

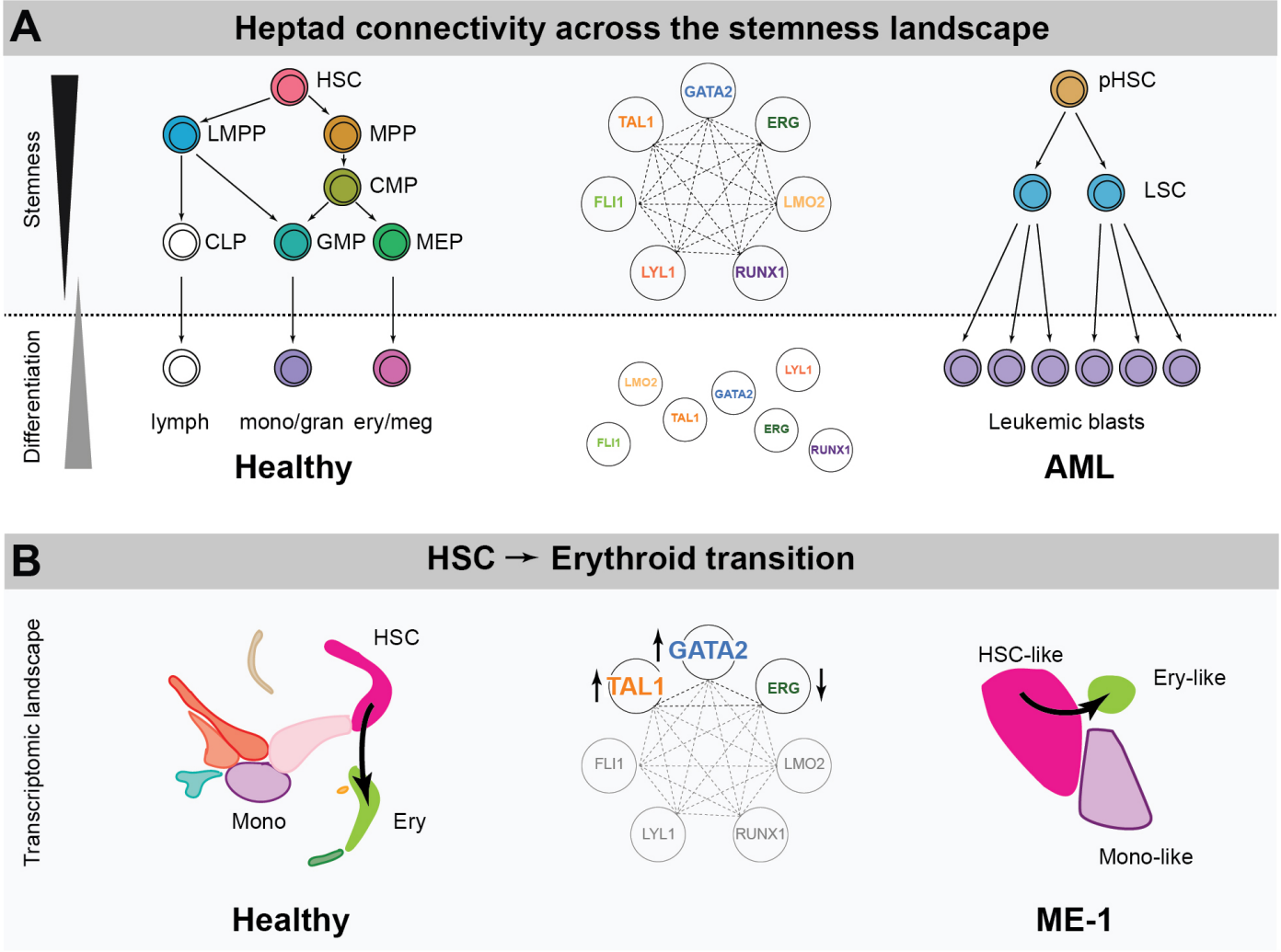
A
Heptad connectivity across the stemness landscape
Stemness
Differentiation
HSC
LMPP
MPP
CMP
CLP
GMP
MEP
lymph
mono/gran
ery/meg
Healthy
GATA2
TAL1
ERG
FLI1
LMO2
LYL1
RUNX1
pHSC
LSC
Leukemic blasts
AML
B
HSC → Erythroid transition
Transcriptomic landscape
HSC
Mono
Ery
Healthy
HSC-like
Ery-like
Mono-like
ME-1

Figure 6

Disruption of a GATA2, TAL1, ERG regulatory circuit promotes erythroid transition in healthy and leukemic stem cells.

Running title: GATA2, TAL1 and ERG regulate erythroid transition.

Julie A. I Thoms[1*], Peter Truong[2], Shruthi Subramanian[2], Kathy Knezevic[2], Gregory Harvey[2], Yizhou Huang[2,3], Janith A. Seneviratne[4], Daniel R. Carter[3,4], Swapna Joshi[2], Joanna Skhinas[2], Diego Chacon[3], Anushi Shah[2], Ineke de Jong[5], Dominik Beck[2,3], Berthold Göttgens[6], Jonas Larsson[5], Jason W. H. Wong[7], Fabio Zanini[2,8, #, *], and John E. Pimanda[1,2,9,#,*]

[1]*School of Medical Sciences and Adult Cancer Program, Faculty of Medicine, UNSW Sydney, NSW 2052, Australia.*

[2]*Prince of Wales Clinical School and Adult Cancer Program, Faculty of Medicine, UNSW Sydney, NSW 2052, Australia.*

[3]*School of Biomedical Engineering, University of Technology Sydney, NSW 2007, Australia.*

[4]*Children's Cancer Institute Australia for Medical Research, Lowy Cancer Research Centre, UNSW Sydney, Kensington, New South Wales, Australia.*

[5]*Division of Molecular Medicine and Gene Therapy, Lund Stem Cell Center, Lund University, SE-22100, Lund, Sweden.*

[6] *Wellcome and MRC Cambridge Stem Cell Institute, Cambridge, United Kingdom.*

[7]*School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region*

[8]*Garvan-Weizmann Centre for Cellular Genomics, Garvan Institute of Medical Research, Sydney, Australia*

[9]*Department of Haematology, Prince of Wales Hospital, Randwick NSW 2031, Australia.*

[#]*Equal contributions,*

* corresponding authors: Julie Thoms- j.thoms@unsw.edu.au , Fabio Zanini- fabio.zanini@unsw.edu.au , John Pimanda- jpimanda@unsw.edu.au ; Level 2, Lowy Cancer Research Centre, UNSW Sydney, NSW, Australia,

P: +612 9385 2527 F: +612 9385 1510

Word count (abstract): 246, Word count (text): 4398, Figure count: 6 main, 8 supplemental, Table count: 10 supplemental, Reference count: 95

Keywords: gene regulation, differentiation, transcriptional networks, haematopoiesis

Key points (2, 140 characters each including spaces)

* Chromatin accessibility patterns at key heptad regulatory elements can predict cell identity in healthy progenitors and leukemic cells.

* A sub-circuit comprised of *GATA2*, *TAL1*, and *ERG* regulates the stem cell to erythroid transition in both healthy and leukemic cells.

ABSTRACT

Changes in gene regulation and expression govern orderly transitions from hematopoietic stem cells to terminally differentiated blood cell types. These transitions are disrupted during leukemic transformation but knowledge of the gene regulatory changes underpinning this process is elusive. We hypothesised that identifying core gene regulatory networks in healthy hematopoietic and leukemic cells could provide insights into network alterations that perturb cell state transitions. A heptad of transcription factors (LYL1, TAL1, LMO2, FLI1, ERG, GATA2, RUNX1) bind key hematopoietic genes in human CD34+ haematopoietic stem and progenitor cells (HSPCs) and have prognostic significance in acute myeloid leukemia (AML). These factors also form a densely interconnected circuit by binding combinatorially at their own, and each other's, regulatory elements. However, their mutual regulation during normal haematopoiesis and in AML cells, and how perturbation of their expression levels influences cell fate decisions remains unclear. Here, we integrated bulk and single cell data and found that the fully connected heptad circuit identified in healthy HSPCs persists with only minor alterations in AML, and that chromatin accessibility at key heptad regulatory elements was predictive of cell identity in both healthy progenitors and in leukemic cells. The heptad factors GATA2, TAL1 and ERG formed an integrated sub-circuit that regulates stem cell to erythroid transition in both healthy and leukemic cells. Components of this triad could be manipulated to facilitate erythroid transition providing a proof of concept that such regulatory circuits could be harnessed to promote specific cell type transitions and overcome dysregulated haematopoiesis.

INTRODUCTION

Haematopoietic stem cells (HSCs) reside in the bone marrow niche where they are mostly quiescent but retain the capacity to self-renew and replace terminal blood cell types throughout life[1]. Haematopoiesis is a hierarchical process with HSCs at the apex giving rise to a range of progenitor cells with increasing lineage restriction[1]. Although single cell transcriptomic data suggest a continuous differentiation process[2-7], relatively pure progenitor populations corresponding to intermediate differentiation stages can be prospectively isolated based on surface marker expression[3]. Cell type transitions are controlled by cell intrinsic and extrinsic factors, and loss of control can lead to inappropriate proliferation and leukemic transformation[8-13].

Acute myeloid leukemia (AML) is characterised by an abundance of relatively undifferentiated cells (blasts) of the myeloid lineage[14]. AMLs likely originate in the earliest HSC compartments or acquire stem-cell-like transcriptional programs during leukemic transformation[15-19]. Although blast cells can comprise the bulk of the AML population, self-renewal is restricted to a smaller population of leukemic stem cells (LSCs) which can recapitulate the disease after ablation of the blast population[20-22]. LSCs drive relapse[23], potentially because they possess stem cell transcriptional programs[24,25]. Thus, AML induces a parallel hierarchy of malignant cell types with LSCs at the top[26]. Therapies that induce LSC differentiation by targeting mutant proteins that block differentiation are effective but limited to a minority of AMLs[27-31].

AML is a heterogenous disease with numerous driver mutations[14,32-34], many of which converge on corruption of the transcriptional networks that control normal haematopoiesis[13,35-37]. Transcriptional networks coordinate gene regulation and play a key

role in establishing and maintaining cell identity throughout the life of an organism[12,38]. Such networks are cell type specific, and therefore need to be rewired during embryonic development and differentiation, while disruption can lead to oncogenic transformation[8-13]. Indeed, transcriptional networks are altered across AMLs with a wide spectrum of mutational origins, such that AML cells assume a new epigenetic identity distinct from any normal blood cell type[35]. Furthermore, epigenetic rewiring is increasingly being recognised as a non-genetic cause of treatment resistance[39-41]. However, the specific molecular mechanisms underlying disruption of transcriptional networks in AML, and whether these can be therapeutically targeted, remain unknown.

We and others have previously described seven transcriptional regulators (heptad; LYL1, TAL1, LMO2, FLI1, ERG, GATA2, RUNX1) which bind to key haematopoietic genes in normal human CD34+ haematopoietic stem and progenitor cells (HSPCs) and in AML[42-44]. Heptad factors also bind combinatorially at their own, and each other's, regulatory elements, forming a densely interconnected circuit that plays a role in maintaining the stem cell state [42,44]. The heptad circuit appears to be established at the haemogenic endothelium stage of blood development[45], and over-expression of all seven factors in a mouse *in vitro* differentiation system led to increased production of pre-HSPCs with capacity for multilineage differentiation [46]. All seven factors are key haematopoietic regulators, and mutation or dysregulation is commonly associated with haematological or other malignancies[32,47-50]. Furthermore, the heptad circuit is maintained or reactivated in AML[43,51-53], and heptad expression is predictive of patient outcome[43]. However, heptad circuitry and function have primarily been established using bulk ChIPseq experiments in heterogenous cell populations (i.e. HSPCs) which may obscure underlying sub-circuits or relationships that only exist in specific cell types/cellular contexts. Thus, key questions remain about the

precise roles of the heptad throughout normal and leukemic haematopoiesis, including whether all seven factors act together in single cells, and whether heptad TFs contribute to cell fate decisions as well as maintaining stemness.

Here we integrate bulk and single cell data in normal human HSPCs and leukemic cells and find that chromatin conformation at key heptad regulatory elements is predictive of cell identity in normal and leukemic progenitors. The interconnected heptad circuit identified in normal HSPCs persists in AML, but single cell transcriptomics suggest that specific heptad sub-circuits exist in individual cells and play a key role in determining differentiation trajectories as cells exit the stem cell state.

METHODS

Supplementary Methods detail standard techniques.

NGS data generation/processing

Chromatin immunoprecipitation (ChIP) was performed as described[43] (antibodies in Table S1). Library construction/sequencing was performed by BGI Genomics (China) or Novogene (Hong Kong). Single cell RNA sequencing (scRNAseq) used the 10X Genomics pipeline. Aligned **s**equencing data was displayed in BigWig format, and read counts covering enhancers (Table S2) extracted using deepTools pyBigWig[54] and plotted.

Replicate ATACseq counts were added. Profiles were encoded as unit vectors by dividing by total counts across all heptad peaks. Cityblock distances on the multidimensional unit sphere between each sample and each average profile were used to compute the heatmap and predict cell types.

scRNAseq Analysis

Analysis for Figures 1, 4 is at https://github.com/iosonofabio/heptad_paper. Healthy hematopoietic cells data was downloaded as described https://github.com/dpeerlab/Palantir/blob/master/README.md, Rep1. Embedding coordinates, colours, cluster metadata, and smoothed counts data were extracted from the h5ad file and plotted using singlet (https://github.com/iosonofabio/singlet).

Count and metadata tables from CellRanger (10X Genomics) were converted to loom format (http://loompy.org/) and normalised to "counts per ten thousand (uniquely mapped) reads". The symmetric correlation matrix was ordered by hierarchical (average linkage)

clustering on L2 distance with optimal leaf ordering. Conditional distributions of gene expression were computed via quantiles followed by kernel density estimate in logarithmic space.

Palantir data were subsampled to 40 cells/type. northstar's subsample method[55] was used to infer cell states within ME-1 guided by Palantir data[6]. For graph construction, 10 external (non-mutual) neighbours were allowed to compensate for the fact that ME-1 cells are quite distant from actual hematopoietic cells. RNA velocity[56] was computed using scVelo[57] and projected onto northstar's embedding. Gene expression was plotted in the same embedding after iterative nearest-neighbour smoothing. For predicting ME-1 cell state, we trained a random forest classifier using scikit-learn and evaluated its performance via train/test splits.

Data sharing

Table S3 shows public datasets. New data is deposited under accession GSE158797. Code is available from https://github.com/iosonofabio/heptad_paper.

RESULTS

Heptad expression during haematopoiesis

To understand heptad expression patterns during haematopoiesis we interrogated existing scRNAseq data (Palantir) from bone marrow cells[6] (Figure 1A). Diverging patterns of heptad transcription factor (TF) expression were observed across developmental time (Figure 1B). All seven TFs are expressed in HSCs, with increasing divergence during differentiation. For example, *GATA2*, *TAL1*, *LYL1*, and *LMO2* are upregulated along the erythroid lineage, while *RUNX1* is upregulated along the granulocytic/monocytic lineage.

Heptad regulatory region accessibility during normal haematopoiesis

Heptad TFs form a densely interconnected circuit in bulk CD34+ HSPCs, with each corresponding gene having regulatory regions bound by most of the heptad[42]. Since heptad expression patterns are heterogeneous in single cells, we asked whether there is evidence for changes in heptad regulation at any of the combinatorially bound regions over developmental time. Although haematopoiesis is a continuum (Figure 1A), functionally defined subpopulations representing various waypoints can be isolated based on cell surface marker expression (Figure 1C). We queried chromatin accessibility data from sorted bone marrow subpopulations[4], focussing on known heptad gene regulatory regions (*LYL1* promoter (P), *TAL1*+40, *LMO2*-25, *FLI1*-16, *ERG*+85, *GATA2*+3.5, *RUNX1*+23[42]). We included two putative regulatory regions; *RUNX1*+141, an intragenic *RUNX1* region that was heptad-bound in HSPCs[42], and *GATA2*-117, a distal regulatory element for *GATA2* that is dysregulated by translocation in the inv(3) AML subtype[58,59]. Strikingly, accessibility patterns differed throughout development with some elements (*FLI1*-15, *ERG*+85, *GATA2*+3.5, *RUNX1*+141) losing accessibility upon exiting the $CD34^+$ progenitor stage, suggesting that heptad connectivity is lost once cells commit to terminal differentiation (Figure 1D). Individual heptad regulatory elements remain accessible in more differentiated cells (*LYL1*P, *LMO2*-25, *RUNX1*+23; monocyte lineage, and *LYL1*P and *TAL1*+40; erythroid lineage) consistent with expression of the related TF in these cells, with some exceptions such as the *LMO2*-25 enhancer, which is inaccessible in erythroid cells, even though *LMO2* is highly expressed, presumably controlled by alternate regulatory regions. The *TAL1*+40 and *GATA2*-117 elements had the most restricted accessibility patterns with both biased toward the erythroid lineage in line with higher expression of *TAL1* and *GATA2* in these cells.

Heptad regulatory region accessibility in AML

The heptad circuit can be active in AML[43,51-53] and heptad expression can predict patient survival[43]. Data from two cohorts of AML cells showed that heptad regulatory regions were accessible in AMLs with diverse molecular lesions[35] (Figure S1A) and in pre-leukemic HSCs, LSCs, and leukemic blasts isolated from the same patient[4] (Figures 1E, S1B). Notably, the *TAL1*+40 enhancer was rarely accessible in AML, and the *GATA2*-117 enhancer varied between patient samples.

Heptad regulatory region accessibility can classify normal and leukemic cells

Genome-wide chromatin accessibility profiles reflect cell identity[4]. Since heptad expression and regulatory region accessibility are heterogenous throughout development, we asked whether the pattern of chromatin accessibility at heptad regulatory regions is sufficient to predict cell type. Using a classifier based on nine regulatory regions, we could correctly identify normal cells across the haematopoietic spectrum (Figure 1F). Furthermore, this classifier could assign a "closest normal" type to AML samples sorted into pre-leukemic HSC (pHSC), LSC, and blast populations (Figure 1G). Consistent with known AML biology, pHSCs were predominantly classified as HSCs or MPPs, LSCs as LMPPs or GMPs, and blasts as more variable cell types. We compared our cell type assignments to published classifications of these samples based on whole genome accessibility patterns[4] and found a high concordance in pHSC and LSC populations (Figures 1H, S1C). Consistent with lost heptad connectivity in more differentiated cells, the heptad-based classifier had reduced concordance with genome-wide classification in blast populations. Overall, our analysis indicates that heptad expression and accessibility are associated with cell identity in healthy haematopoietic progenitors and leukemic cells.

The heptad network persists in AML, with altered connectivity.

We extended our analysis and asked which heptad TFs were bound at each regulatory region in normal and AML contexts, looking first at heptad binding patterns at the nine regulatory regions in $CD34^+$ HSPCs[42] (Figure 2A, *left*, Figure S2). Combinatorial binding was observed, with LYL1, FLI1, GATA2, and RUNX1 bound at all regions, and *FLI1*, *ERG*, *GATA2*, and *RUNX1* each having at least one regulatory element bound by all seven TFs. Binding patterns were then used to infer the connectivity map of heptad autoregulation in HSPCs (Figure 2A, *right*).

We next compared heptad connectivity in two AML cell lines, ME-1, and KG-1. AML cell lines recapitulate properties of primary AML cells[60] and can be experimentally manipulated. ME-1 and KG-1 cells express all seven heptad genes, although the pattern of individual TF expression varies both between cell lines and compared to HSPCs (Figure S3). Consistent with primary AML accessibility, heptad ChIPseq in ME-1 (Figures 2B, S4) and KG-1 (Figure 2C, S5) revealed that the densely interconnected circuit observed in HSPCs persists in AML cells, although the precise pattern of connectivity varies. For example, both ME-1 and KG-1 have prominent binding peaks at *LYL1*P, while at *TAL1*+40, ME-1 and KG-1 had fewer called peaks (4/7 and 2/7 respectively) than HSPCs (5/7), and these were generally small. Overall, heptad TFs remain highly connected in both AML cell lines, albeit with somewhat different circuit structures compared to HSPCs. Expression levels of individual TFs in HSPCs and AML cell lines were broadly in keeping with the number and binding intensities of TFs at the cognate regulatory element (Figure 2, S3), except for LMO2 which had similar numbers and sizes of ChIPseq peaks across all cell types but was highly expressed in HSPCs.

Heptad regulatory elements require ETS and GATA motifs

Having shown that heptad binding at regulatory regions persists in AML, we wanted to understand the role of specific TF binding motifs within these regulatory regions. *Cis*-regulatory elements integrate signals from multiple TFs which bind to specific DNA sequences, with direct binding occurring at consensus binding motifs. The heptad TFs belong to four broad classes of TFs with different consensus binding motifs – E-box (CANNTG, bound directly by LYL1 and TAL1 and indirectly by LMO2), ETS (GGAW, bound by FLI1 and ERG), GATA (GATA, bound by GATA2) and RUNX (TGYGGT, bound by RUNX1). To identify consensus motifs likely to correspond to TF binding sites, we performed multiple sequence alignments using human, mouse, dog, and opossum genomes (Figure 3A). All regulatory elements contained conserved ETS and GATA motifs, while 7/9 contained a conserved E-Box motif and 6/9 a conserved RUNX motif. We mutated all conserved instances of each binding motif class (Table S4) and tested in luciferase reporter constructs in KG-1 and ME-1 cells.

Deletion of ETS consensus motifs was universally deleterious, leading to significant loss of activity for all elements tested (Figure 3B). Deletion of GATA consensus motifs had a significant negative impact for all regions in at least one cell line. Deletion of E-box or RUNX motifs reduced luciferase reporter activity, however the effect was generally small compared to deletion of ETS or GATA motifs, and in one case (*LMO2*-25) deletion of the RUNX motif led to slightly increased activity. Overall, regulatory region activity was impaired by loss of any class of TF binding motif, with loss of ETS or GATA motifs dominating. Two WT reporter constructs, *TAL1*+40 and *RUNX1*+141, showed minimal activity in one or both cell lines (Figure 3C), and were excluded from mutation analysis. Consistent with their activity, *TAL1*+40 had few heptad TF binding inputs in either cell line,

and *RUNX1*+141, which was active in ME-1 but not KG-1, had fewer inputs in KG-1 than in ME-1.

Single cell transcriptomics reveal key regulators of the HSC – erythroid transition

Altered enhancer activity reads out as gene expression changes. Encouraged by our results indicating that removing specific consensus motifs altered activity of heptad regulatory regions, we proceeded to scRNAseq analysis of heptad expression in ME-1 cells which are amenable to downstream perturbation. We quantified heptad heterogeneity and observed that for both high (e.g., *LYL1*) and low (e.g., *ERG*) expressed genes heterogeneity across the ME-1 population spanned an order of magnitude (Figure 4A). Furthermore, the highest gene expression (*LYL1*) corresponded to the highest heptad binding at an associated regulatory region, while lower gene expression (*TAL1* and *GATA2*) corresponded to lower heptad binding at their associated regulatory regions (Figure 2B).

We next looked for pairwise expression correlations between TFs and found *GATA2* was positively correlated with *TAL1*, and negatively correlated with *ERG* and *LMO2* (Figure 4B). Because correlation measures are insensitive to extreme phenotypes, we performed complementary analysis to evaluate whether this effect is also seen at the extreme of the distribution and plotted conditional gene expression distribution in the bottom and top quantiles of expressors of *GATA2* (Figure 4C). Given the observed heterogeneity in heptad expression in ME-1 cells, and the strong association between heptad regulation and cell type we asked whether we could identify subpopulations within the ME-1 scRNAseq data. A canonical unsupervised clustering approach based on overdispersed features did not result in distinct biological patterns beyond cell cycle, as expected from a cell line. We reasoned that a more sophisticated feature selection together with soft guidance from healthy marrow data

could reveal additional hidden heterogeneity. We therefore switched from unsupervised clustering to northstar, a semi-supervised clustering algorithm that leverages information from training data to channel the axes of heterogeneity during feature selection, graph construction, and cell community detection[55]. Using healthy marrow transcriptomes[6] (Figure 1A) as training data, this analysis revealed two major subpopulations, HSC-like (pink) and Mono-precursor-like (purple, 1136 and 277 out of 1489 cells respectively) plus a minor population that was more similar to Ery-precursor cells (lime, 47 out of 1489 cells) and two small groups of cells resembling Megakaryocytes (18 cells) and Monocytes (11 cells, Figure 4D). RNA velocity analysis[56] (Figure 4D *arrows*) revealed a major trajectory along the HSC-Mono-precursor axis, and an alternate trajectory connecting the HSCs to the Ery-precursor population. This flow diagram (independent of northstar clustering) confirmed population structure reminiscent of healthy haematopoiesis (Figure 4D *inset*). Primary AML cells also have population structures resembling normal haematopoiesis[61] and have differential heptad expression between subpopulations (Figure S6A). We projected expression levels of the four previously identified genes on embedded cell plots (Figure 4E), and consistent with our correlation data and known biological functions, *GATA2* and *TAL1* expression were enriched in the Ery-precursor population. Conversely, *ERG* and *LMO2* expression were enriched in the HSC-like and Mono-precursor-like populations. We then computed the fold expression change in heptad genes between HSC and Ery-precursor cells in both ME-1 and normal BM cells (Figures 4F, S6B, S6C, Tables S5, S6). In ME-1 cells, *ERG* expression was reduced (0.6x) and *GATA2* and *TAL1* expression increased (11x and 3.5x respectively) in Ery-precursor cells (Figure 4F *left*). We observed a similar pattern in healthy cells, although *FLI1*, *RUNX1*, and *LMO2* also showed expression changes in this context (Figure 4F *right*).

To better understand how heptad TFs influence cell-specific gene expression we interrogated TF binding in bulk HSPCs. As these cells are a mixture of progenitor types, we focussed on ATACseq peaks uniquely accessible in HSCs or MEPs (Figure S7, Table S7). *ERG*, *FLI1* and *RUNX1* had higher expression in HSCs compared to Ery-precursors and showed higher average binding at HSC-unique peaks, while *GATA2*, *TAL1*, and *LYL1* were more highly expressed in Ery-precursors but had similar average binding at both MEP- and HSC-unique peaks (Figure S7). *LMO2* had higher expression in Ery-precursors, but higher binding at HSC-unique peaks. TFs bind DNA directly via their cognate binding motifs, or indirectly via protein-protein interactions. HSC-unique peaks were highly enriched for ETS motifs (Table S8, significance value (sv) 5.50E-171), and enriched for RUNX motifs (Table S8, sv 5.70E-08), consistent with higher ERG, FLI1, and RUNX1 binding at these peaks. MEP-unique peaks were bound by GATA2 and highly enriched for GATA motifs (Table S8, sv 3.20E-111). GATA2 was also bound at HSC-unique peaks, although GATA motifs were enriched in only a minor fraction of HSC-unique peaks (Table S8, 33/7396, sv 3.10E-02), suggesting that GATA2 binding at these sites may be mediated by interactions with other transcription factors rather than direct DNA binding.

Finally, we asked whether heptad expression was sufficient to classify ME-1 cells as HSC-like or Ery-Precursor-like (Figure 4G). Using a random forest classifier based on Palantir data, we found heptad expression was able to correctly classify cells with high accuracy (area under ROC = 0.80), and that *GATA*2 expression was the best performing gene in terms of model accuracy (area under ROC = 0.84).

Direct manipulation of GATA2 and ERG promotes erythroid trajectory.

We then evaluated effects of perturbing heptad factors on i) expression of other heptad factors, ii) global transcriptome of perturbed cells, and iii) cell function. Specifically, we predicted that high levels of *GATA2* or *TAL1* and low levels of *ERG* would promote transition along the HSC-Ery-precursor axis (Figure 5A). We first knocked down key heptad genes in ME-1 cells (Figure S8A) and measured the response of other heptad genes. *GATA2* knockdown led to decrease of *TAL1* and most other heptad genes, except for *ERG* which was unaffected by *GATA2* knockdown (Figure 5B, *left*). Similarly, *TAL1* knockdown led to decreased *GATA2* and most other heptad genes except for *ERG* (Figure 5B, *centre*). Conversely, *ERG* knockdown led to decreased *LMO2* expression, but increased expression of *GATA2*, *FLI1*, and *TAL1* (Figure 5B, *right*). *RUNX1* expression showed inconsistent changes, possibly due to dysregulation via translocation of its essential binding partner *CBFβ* in ME-1 cells[62]. Similar results were observed using additional shRNAs targeting *GATA2* or *ERG* (Figure S8B). Heptad gene expression also changed following knockdown of *GATA2*, *TAL1*, or *ERG* in two additional AML cell lines (Figure S8C, S8D), although response patterns varied between cell lines, likely reflecting the unique cell subpopulations in each.

Since the bulk of ME-1 cells were assigned as HSC-like, we reasoned that *ERG* knockdown, or *GATA2* overexpression, might alter their trajectory away from the HSC-like and towards the Ery-precursor-like state. *ERG* knockdown reduced ME-1 colony formation in methylcellulose (Figure S8E), consistent with a shift away from the HSC-like state. We also analysed RNAseq data from *GATA2* over-expression in ME-1 cells[63] and found that increased *GATA2* led to increased *TAL1* and *RUNX1*, and reduced *ERG* and *LMO2*, similar to expression changes between Ery-precursor-like and HSC-like ME-1 cells (Figure 5C, *left*, compare to Figure 4F, *left*). GSEA analysis was used to compare GATA2 driven changes in global gene expression to expression differences between Ery-precursors and HSCs.

Globally, genes that were high in Ery-precursors tended to increase following GATA2 overexpression, while genes that were low in Ery-precursors tended to decrease (Figure 5C, *right*). *ERG* overexpression in HSPCs promotes progenitor expansion[64], and we have now shown that *ERG* expression is reduced across the HSC to Ery-precursor boundary in normal BM and ME-1 (Figure 4F). Furthermore, an independent method using scRNAseq landscapes as references predicts that perturbing *ERG* in mouse or human LMPPs would push cells towards an erythroid fate[65]. We therefore asked whether *ERG* knockdown in HSPCs promoted an Ery-progenitor phenotype. *ERG* knockdown led to downregulation of *FLI1*, *LYL1*, and *LMO2*, and upregulation of *GATA2* and *TAL1* (Figure 5D, *left*), similar to expression changes across the HSC-Ery-progenitor transition in Palantir data (Figure 4F, *right*). GSEA analysis was used to compare *ERG* knockdown driven changes in global gene expression to expression differences between Ery-precursors and HSCs. Globally, genes that were high in Ery-precursors tended to increase following ERG knockdown, while genes that were low in Ery-precursors tended to decrease (Figure 5D, *right*). To evaluate functional consequences of *ERG* knockdown in HSPCs (Figure 5E) we measured colony forming capacity and found that cells with reduced *ERG* expression were skewed towards erythroid colony formation (Figure 5F). Together, the perturbation data supports the notion that heptad genes, and in particular the triplet *GATA2*, *TAL1*, and *ERG*, form a functionally relevant interconnected network and play a key role in regulating cell state transitions in healthy blood and leukemic cells.

DISCUSSION

Gene regulatory networks control cell fate decisions in development and disease. We focused on heptad transcription factors and identified parallel phenotypes between healthy haematopoiesis and leukemic cells spanning single cell gene expression, chromatin state, and enhancer use (Figure 6A). Our data suggest that GATA2, TAL1, and ERG constitute a heptad sub-circuit that regulates stem cell to erythroid transition in healthy blood and leukemia (Figure 6B).

Insights into enhancer biology

Genome-wide chromatin state can be used to classify cell types[4]. We show that chromatin accessibility at only nine heptad enhancers could classify all early stages of haematopoiesis and subpopulations of AML cells. While the transcriptional network determining haematopoietic cell fate decisions undoubtedly contains additional enhancers, the heptad enhancers studied here give significant insight into the transcriptional control of blood cell identity. Most heptad enhancers were accessible in HSPCs and became selectively inaccessible at terminal differentiation, though exceptions were observed. We found the *GATA2*-117 (mice: *Gata2*-77) enhancer was open only in CMPs and MEPs, suggesting a central role for this enhancer in erythroid transition and confirming previous murine models, where its deletion blocked erythroid and megakaryocytic differentiation[66].

This enhancer has been previously studied in inv(3) AML where it is translocated close to oncogene *MECOM*/EVI1 leading to increased EVI1 and decreased *GATA2* expression[58,59]. We found this enhancer was accessible in a subset of leukemic cells, and strongly heptad-bound in both AML cell lines compared to HSPCs. In our reporter assays *GATA2*-117 also drove more luciferase activity than *GATA2*+3.5, the other *GATA2* regulatory

element. Thus, even in its normal genomic context *GATA2*-117 may play a role in driving *GATA2* expression in AML. Unlike *GATA2*-117, the *ERG*+85 enhancer was open in all HSPC subsets and across AML subtypes (Figure S1A). This enhancer has been linked to AML prognosis[43] and used to identify LSCs within bulk AML populations[67,68]. Enhancers are replete with sequence motifs enabling binding of distinct TF families, either directly to DNA or indirectly via protein scaffolding, as observed for LMO2[69,70] and RUNX1[42,44]. Here, we showed that evolutionarily conserved heptad enhancers rely heavily on ETS and GATA motifs, in agreement with previous reports that ETS-ETS-GATA motifs were enriched at blood enhancers[71].

Regulation of cell fate transitions by GATA2, TAL1 and ERG

Combinatorial binding of TFs is a key component of cell fate transitions[38]. We identify a triad of TFs-GATA2, TAL1, and ERG, whereby high *GATA2* and *TAL1*, and low *ERG* expression biased fate decisions towards the erythroid lineage in both HSPCs and ME-1 leukemic cells. A similar circuit, comprised of *GATA2*, *TAL1*, and *FLI1* (an ETS TF closely related to ERG) has been previously reported during embryonic HSC specification[72], while *GATA1*, *TAL1* and KLF1 form a sub-circuit in erythroid cells[73]. Indeed, recycling of regulatory modules is a key feature of developmental networks[38], underlining the utility of cell classification strategies such as northstar[55].

Each member of this triad is known to play complex roles in healthy blood and leukemia development. GATA2 controls blood cell emergence in the embryonic aorta[74], and is required for HSC maintenance[75]. Germline loss of function mutations in *GATA2* predisposes to MDS and AML[76] and high *GATA2* expression is associated with poor prognosis in AML patients[77]. *TAL1* is also required for embryonic blood formation[48,78] and

drives erythroid and megakaryocytic differentiation programs[79] but is dispensible for HSC maintenance[48,80,81]. However, dysregulation of TAL-1 is associated with T-ALL[48]. *ERG* is not required for HSC specification or differentiation but promotes HSC maintenance by restricting differentiation[82,83]. High *ERG* expression is a poor prognostic marker for AML[49,84-86] and is leukemogenic in mouse models[87-90], although its role in human leukemia is more subtle[64].

Clinical Implications

Therapeutic approaches to AML which force LSCs to differentiate have been sought[91]. Although TFs are relatively difficult drug targets, small molecules upregulating CEBPA[92,93] or downregulating PU.1[94] and RUNX1[95] have been developed. Regulatory circuits such as the GATA2-TAL1-ERG triad described here may provide a conceptual framework within which to develop such therapies. A first approach would be to alter TF expression directly, as upregulating *GATA2* or downregulating *ERG* promotes erythroid differentiation. However, population structure of malignant cells within primary AML varies between patients and different leukemias may be primed towards specific differentiation pathways[61]. As such, *ERG* perturbation is especially promising as this TF appears to preserve the progenitor state rather than bias towards a particular fate, and knockdown may favour exit from the stem cell state across a range of primary AMLs. A second approach would be to focus on transcriptional regulators of these TFs. USP9X, a deubiquitinase that regulates ERG stability[96] and is positively regulated by ERG in a feed forward loop is one such candidate[67]. A third approach would be to focus on specific enhancers such as *GATA2*-117, which is inaccessible in normal HSCs but open in the transitional progenitor states characteristic of AML, enabling preferential cytotoxicity in leukemic cells. Overall, a deeper understanding of

heptad regulatory circuits and their roles in maintaining and exiting normal and leukemic stem cell states can help shape novel, data-based approaches to innovative cancer therapies.

ACKNOWLEDGEMENTS

The authors thank the staff and donors of the Sydney Cord Blood Bank for providing cord bloods for research. Some of the data presented in this work was acquired by personnel and/or instruments at the Mark Wainwright Analytical Centre (MWAC) of UNSW Sydney, which is in part funded by the Research Infrastructure Programme of UNSW. The authors acknowledge the following funding support: JT was supported by the Anthony Rothe Memorial Trust; PT is supported by a University International Postgraduate Award from UNSW Sydney and Translational Cancer Research Network - a Translational Cancer Research Centre funded by the Cancer Institute NSW. SS is supported by an International Postgraduate Student scholarship from UNSW Sydney and the Prince of Wales Clinical School and Translational Cancer Research Network - a Translational Cancer Research Centre funded by the Cancer Institute NSW; DB is supported by a Peter Doherty Fellowship from the National Health and Medical Research Council of Australia (APP1073768), a Cancer Institute NSW Early Career Fellowship, the Anthony Rothe Memorial Trust, and Gilead Sciences; BG is supported by a Wellcome Investigator award (206328/Z/17/Z); JEP is supported by project grants from the National Health and Medical Research Council of Australia (APP1042934, APP1102589, APP1008515), a translational program grant from the Leukemia Lymphoma Society (LLS)-Snowdome Foundation-Leukaemia Foundation, project funds from the Translational Cancer Research Network - a Translational Cancer Research Centre funded by the Cancer Institute NSW, Anthony Rothe Memorial Trust, and philanthropic funding from Christina's Light.

AUTHORSHIP CONTRIBUTIONS

J.A.I.T., P.T., S.S., K.K., G.H., Y.H., J.S., D.R.C., S.J., and J.S. performed research and analysed data. D.C., A.S., D.B., and J.W.H.W. analysed data. I.dJ. and J.L. provided key

reagents. B.G. and J.W.H.W. discussed and interpreted data. J.A.I.T., F.Z., and J.P. conceived the study and wrote the paper.

CONFLICT OF INTEREST DISCLOSURES

The authors report no financial conflicts.

REFERENCES

1. Doulatov S, Notta F, Laurenti E, Dick JE. Hematopoiesis: a human perspective. *Cell Stem Cell*. 2012;10(2):120-136.
2. Velten L, Haas SF, Raffel S, et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol*. 2017;19(4):271-281.
3. Buenrostro JD, Corces MR, Lareau CA, et al. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell*. 2018;173(6):1535-1548 e1516.
4. Corces MR, Buenrostro JD, Wu B, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet*. 2016;48(10):1193-1203.
5. Karamitros D, Stoilova B, Aboukhalil Z, et al. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat Immunol*. 2018;19(1):85-97.
6. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol*. 2019;37(4):451-460.
7. Watcham S, Kucinski I, Gottgens B. New Insights into Haematopoietic Differentiation Landscapes from scRNA-seq. *Blood*. 2019.
8. Pimanda JE, Gottgens B. Gene regulatory networks governing haematopoietic stem cell development and identity. *Int J Dev Biol*. 2010;54(6-7):1201-1211.
9. Sive JI, Gottgens B. Transcriptional network control of normal and leukaemic haematopoiesis. *Exp Cell Res*. 2014;329(2):255-264.
10. Enver T, Pera M, Peterson C, Andrews PW. Stem cell states, fates, and the rules of attraction. *Cell Stem Cell*. 2009;4(5):387-397.
11. Moris N, Pina C, Arias AM. Transition states and cell fate decisions in epigenetic landscapes. *Nat Rev Genet*. 2016;17(11):693-703.
12. Wilkinson AC, Nakauchi H, Gottgens B. Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity. *Cell Syst*. 2017;5(4):319-331.
13. Thoms JAI, Beck D, Pimanda JE. Transcriptional networks in acute myeloid leukemia. *Genes Chromosomes Cancer*. 2019;58(12):859-874.
14. Dohner H, Weisdorf DJ, Bloomfield CD. Acute Myeloid Leukemia. *N Engl J Med*. 2015;373(12):1136-1152.
15. Horton SJ, Huntly BJ. Recent advances in acute myeloid leukemia stem cell biology. *Haematologica*. 2012;97(7):966-974.
16. Jan M, Snyder TM, Corces-Zimmerman MR, et al. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci Transl Med*. 2012;4(149):149ra118.
17. Shlush LI, Zandi S, Mitchell A, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*. 2014;506(7488):328-333.
18. Basilico S, Gottgens B. Dysregulation of haematopoietic stem cell regulatory programs in acute myeloid leukaemia. *J Mol Med (Berl)*. 2017;95(7):719-727.
19. Corces-Zimmerman MR, Hong WJ, Weissman IL, Medeiros BC, Majeti R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc Natl Acad Sci U S A*. 2014;111(7):2548-2553.
20. Lapidot T, Sirard C, Vormoor J, et al. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature*. 1994;367(6464):645-648.
21. Goardon N, Marchi E, Atzberger A, et al. Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer Cell*. 2011;19(1):138-152.
22. Sarry JE, Murphy K, Perry R, et al. Human acute myelogenous leukemia stem cells are rare and heterogeneous when assayed in NOD/SCID/IL2Rgammac-deficient mice. *J Clin Invest*. 2011;121(1):384-395.

23. Shlush LI, Mitchell A, Heisler L, et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature*. 2017;547(7661):104-108.
24. Eppert K, Takenaka K, Lechman ER, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat Med*. 2011;17(9):1086-1093.
25. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. *JAMA*. 2010;304(24):2706-2715.
26. Bonnet D, Dick JE. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat Med*. 1997;3(7):730-737.
27. Sanz MA, Grimwade D, Tallman MS, et al. Management of acute promyelocytic leukemia: recommendations from an expert panel on behalf of the European LeukemiaNet. *Blood*. 2009;113(9):1875-1891.
28. DiNardo CD, Stein EM, de Botton S, et al. Durable Remissions with Ivosidenib in IDH1-Mutated Relapsed or Refractory AML. *N Engl J Med*. 2018;378(25):2386-2398.
29. Stein EM, DiNardo CD, Pollyea DA, et al. Enasidenib in mutant IDH2 relapsed or refractory acute myeloid leukemia. *Blood*. 2017;130(6):722-731.
30. Hansen E, Quivoron C, Straley K, et al. AG-120, an Oral, Selective, First-in-Class, Potent Inhibitor of Mutant IDH1, Reduces Intracellular 2HG and Induces Cellular Differentiation in TF-1 R132H Cells and Primary Human IDH1 Mutant AML Patient Samples Treated Ex Vivo. *Blood*. 2014;124(21):3734-3734.
31. Popovici-Muller J, Lemieux RM, Artin E, et al. Discovery of AG-120 (Ivosidenib): A First-in-Class Mutant IDH1 Inhibitor for the Treatment of IDH1 Mutant Cancers. *ACS Medicinal Chemistry Letters*. 2018;9(4):300-305.
32. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med*. 2016;374(23):2209-2221.
33. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391-2405.
34. Cancer Genome Atlas Research N, Ley TJ, Miller C, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*. 2013;368(22):2059-2074.
35. Assi SA, Imperato MR, Coleman DJL, et al. Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nat Genet*. 2019;51(1):151-162.
36. Yi G, Wierenga ATJ, Petraglia F, et al. Chromatin-Based Classification of Genetically Heterogeneous AMLs into Two Distinct Subtypes with Diverse Stemness Phenotypes. *Cell Rep*. 2019;26(4):1059-1069 e1056.
37. McKeown MR, Corces MR, Eaton ML, et al. Superenhancer Analysis Defines Novel Epigenomic Subtypes of Non-APL AML, Including an RARalpha Dependency Targetable by SY-1425, a Potent and Selective RARalpha Agonist. *Cancer Discov*. 2017;7(10):1136-1153.
38. Davidson EH. Emerging properties of animal gene regulatory networks. *Nature*. 2010;468(7326):911-920.
39. Bell CC, Fennell KA, Chan YC, et al. Targeting enhancer switching overcomes non-genetic drug resistance in acute myeloid leukaemia. *Nat Commun*. 2019;10(1):2723.
40. Fennell KA, Bell CC, Dawson MA. Epigenetic therapies in acute myeloid leukemia: where to from here? *Blood*. 2019;134(22):1891-1901.
41. Guo L, Li J, Zeng H, et al. A combination strategy targeting enhancer plasticity exerts synergistic lethality against BETi-resistant leukemia cells. *Nat Commun*. 2020;11(1):740.
42. Beck D, Thoms JA, Perera D, 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.
43. Diffner E, Beck D, Gudgin E, et al. Activity of a heptad of transcription factors is associated with stem cell programs and clinical outcome in acute myeloid leukemia. *Blood*. 2013;121(12):2289-2300.

44. Wilson NK, Foster SD, Wang X, et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*. 2010;7(4):532-544.
45. Guibentif C, Ronn RE, Boiers C, et al. Single-Cell Analysis Identifies Distinct Stages of Human Endothelial-to-Hematopoietic Transition. *Cell Rep*. 2017;19(1):10-19.
46. Bergiers I, Andrews T, Vargel Bolukbasi O, et al. Single-cell transcriptomics reveals a new dynamical function of transcription factors during embryonic hematopoiesis. *Elife*. 2018;7.
47. Oram SH, Thoms JA, Pridans C, et al. A previously unrecognized promoter of LMO2 forms part of a transcriptional regulatory circuit mediating LMO2 expression in a subset of T-acute lymphoblastic leukaemia patients. *Oncogene*. 2010;29(43):5796-5808.
48. Curtis DJ, Salmon JM, Pimanda JE. Concise review: Blood relatives: formation and regulation of hematopoietic stem cells by the basic helix-loop-helix transcription factors stem cell leukemia and lymphoblastic leukemia-derived sequence 1. *Stem Cells*. 2012;30(6):1053-1058.
49. Marcucci G, Baldus CD, Ruppert AS, et al. Overexpression of the ETS-related gene, ERG, predicts a worse outcome in acute myeloid leukemia with normal karyotype: a Cancer and Leukemia Group B study. *J Clin Oncol*. 2005;23(36):9234-9242.
50. Li Y, Luo H, Liu T, Zacksenhaus E, Ben-David Y. The ets transcription factor Fli-1 in development, cancer and disease. *Oncogene*. 2015;34(16):2022-2031.
51. Mandoli A, Singh AA, Jansen PW, et al. CBFB-MYH11/RUNX1 together with a compendium of hematopoietic regulators, chromatin modifiers and basal transcription factors occupies self-renewal genes in inv(16) acute myeloid leukemia. *Leukemia*. 2014;28(4):770-778.
52. Mandoli A, Singh AA, Prange KHM, et al. The Hematopoietic Transcription Factors RUNX1 and ERG Prevent AML1-ETO Oncogene Overexpression and Onset of the Apoptosis Program in t(8;21) AMLs. *Cell Rep*. 2016;17(8):2087-2100.
53. Sotoca AM, Prange KH, Reijnders B, et al. The oncofusion protein FUS-ERG targets key hematopoietic regulators and modulates the all-trans retinoic acid signaling pathway in t(16;21) acute myeloid leukemia. *Oncogene*. 2016;35(15):1965-1976.
54. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*. 2014;42(Web Server issue):W187-191.
55. Zanini F, Berghuis BA, Jones RC, et al. Northstar enables automatic classification of known and novel cell types from tumor samples. *Sci Rep*. 2020;10(1):15251.
56. La Manno G, Soldatov R, Zeisel A, et al. RNA velocity of single cells. *Nature*. 2018;560(7719):494-498.
57. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol*. 2020.
58. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell*. 2014;25(4):415-427.
59. Groschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369-381.
60. Rucker FG, Sander S, Dohner K, Dohner H, Pollack JR, Bullinger L. Molecular profiling reveals myeloid leukemia cell lines to be faithful model systems characterized by distinct genomic aberrations. *Leukemia*. 2006;20(6):994-1001.
61. van Galen P, Hovestadt V, Wadsworth Ii MH, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell*. 2019;176(6):1265-1281 e1224.
62. Yanagisawa K, Horiuchi T, Fujita S. Establishment and characterization of a new human leukemia cell line derived from M4E0. *Blood*. 1991;78(2):451-457.
63. Yi G, Mandoli A, Jussen L, et al. CBFbeta-MYH11 interferes with megakaryocyte differentiation via modulating a gene program that includes GATA2 and KLF1. *Blood Cancer J*. 2019;9(3):33.

64. Tursky ML, Beck D, Thoms JA, et al. Overexpression of ERG in cord blood progenitors promotes expansion and recapitulates molecular signatures of high ERG leukemias. *Leukemia*. 2015;29(4):819-827.
65. Kucinski I, Wilson NK, Hannah R, et al. Interactions between lineage-associated transcription factors govern haematopoietic progenitor states. *EMBO J*. 2020;39(24):e104983.
66. Johnson KD, Conn DJ, Shishkova E, et al. Constructing and deconstructing GATA2-regulated cell fate programs to establish developmental trajectories. *J Exp Med*. 2020;217(11).
67. Aqaqe N, Yassin M, Yassin AA, et al. An ERG Enhancer-Based Reporter Identifies Leukemia Cells with Elevated Leukemogenic Potential Driven by ERG-USP9X Feed-Forward Regulation. *Cancer Res*. 2019;79(15):3862-3876.
68. Yassin M, Aqaqe N, Yassin AA, et al. A novel method for detecting the cellular stemness state in normal and leukemic human hematopoietic cells can predict disease outcome and drug sensitivity. *Leukemia*. 2019;33(8):2061-2077.
69. Osada H, Grutz G, Axelson H, Forster A, Rabbitts TH. Association of erythroid transcription factors: complexes involving the LIM protein RBTN2 and the zinc-finger protein GATA1. *Proc Natl Acad Sci U S A*. 1995;92(21):9585-9589.
70. Wadman I, Li J, Bash RO, et al. Specific in vivo association between the bHLH and LIM proteins implicated in human T cell leukemia. *EMBO J*. 1994;13(20):4831-4839.
71. Donaldson IJ, Chapman M, Kinston S, et al. Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development. *Hum Mol Genet*. 2005;14(5):595-601.
72. Pimanda JE, Ottersbach K, Knezevic K, et al. Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proc Natl Acad Sci U S A*. 2007;104(45):17692-17697.
73. Wontakal SN, Guo X, Smith C, et al. A core erythroid transcriptional network is repressed by a master regulator of myelo-lymphoid differentiation. *Proc Natl Acad Sci U S A*. 2012;109(10):3832-3837.
74. Eich C, Arlt J, Vink CS, et al. In vivo single cell analysis reveals Gata2 dynamics in cells transitioning to hematopoietic fate. *J Exp Med*. 2018;215(1):233-248.
75. Menendez-Gonzalez JB, Vukovic M, Abdelfattah A, et al. Gata2 as a Crucial Regulator of Stem Cells in Adult Hematopoiesis and Acute Myeloid Leukemia. *Stem Cell Reports*. 2019;13(2):291-306.
76. Hahn CN, Chong CE, Carmichael CL, et al. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat Genet*. 2011;43(10):1012-1017.
77. Vicente C, Vazquez I, Conchillo A, et al. Overexpression of GATA2 predicts an adverse prognosis for patients with acute myeloid leukemia and it is associated with distinct molecular abnormalities. *Leukemia*. 2012;26(3):550-554.
78. Lancrin C, Sroczynska P, Stephenson C, Allen T, Kouskoff V, Lacaud G. The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature*. 2009;457(7231):892-895.
79. Elwood NJ, Zogos H, Pereira DS, Dick JE, Begley CG. Enhanced megakaryocyte and erythroid development from normal human CD34(+) cells: consequence of enforced expression of SCL. *Blood*. 1998;91(10):3756-3765.
80. Mikkola HK, Klintman J, Yang H, 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):547-551.
81. Robertson SM, Kennedy M, Shannon JM, Keller G. A transitional stage in the commitment of mesoderm to hematopoiesis requiring the transcription factor SCL/tal-1. *Development*. 2000;127(11):2447-2459.
82. Taoudi S, Bee T, Hilton A, et al. ERG dependence distinguishes developmental control of hematopoietic stem cell maintenance from hematopoietic specification. *Genes Dev*. 2011;25(3):251-262.

83. Knudsen KJ, Rehn M, Hasemann MS, et al. ERG promotes the maintenance of hematopoietic stem cells by restricting their differentiation. *Genes Dev*. 2015;29(18):1915-1929.
84. Marcucci G, Maharry K, Whitman SP, et al. High expression levels of the ETS-related gene, ERG, predict adverse outcome and improve molecular risk-based classification of cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B Study. *J Clin Oncol*. 2007;25(22):3337-3343.
85. Schwind S, Marcucci G, Maharry K, et al. BAALC and ERG expression levels are associated with outcome and distinct gene and microRNA expression profiles in older patients with de novo cytogenetically normal acute myeloid leukemia: a Cancer and Leukemia Group B study. *Blood*. 2010;116(25):5660-5669.
86. Metzeler KH, Dufour A, Benthaus T, et al. ERG expression is an independent prognostic factor and allows refined risk stratification in cytogenetically normal acute myeloid leukemia: a comprehensive analysis of ERG, MN1, and BAALC transcript levels using oligonucleotide microarrays. *J Clin Oncol*. 2009;27(30):5031-5038.
87. Thoms JA, Birger Y, Foster S, et al. ERG promotes T-acute lymphoblastic leukemia and is transcriptionally regulated in leukemic cells by a stem cell enhancer. *Blood*. 2011;117(26):7079-7089.
88. Goldberg L, Tijssen MR, Birger Y, et al. Genome-scale expression and transcription factor binding profiles reveal therapeutic targets in transgenic ERG myeloid leukemia. *Blood*. 2013;122(15):2694-2703.
89. Carmichael CL, Metcalf D, Henley KJ, et al. Hematopoietic overexpression of the transcription factor Erg induces lymphoid and erythro-megakaryocytic leukemia. *Proc Natl Acad Sci U S A*. 2012;109(38):15437-15442.
90. Salek-Ardakani S, Smooha G, de Boer J, et al. ERG is a megakaryocytic oncogene. *Cancer Res*. 2009;69(11):4665-4673.
91. Nowak D, Stewart D, Koeffler HP. Differentiation therapy of leukemia: 3 decades of development. *Blood*. 2009;113(16):3655-3665.
92. Namasu CY, Katzerke C, Brauer-Hartmann D, et al. ABR, a novel inducer of transcription factor C/EBPalpha, contributes to myeloid differentiation and is a favorable prognostic factor in acute myeloid leukemia. *Oncotarget*. 2017;8(61):103626-103639.
93. Radomska HS, Jernigan F, Nakayama S, et al. A Cell-Based High-Throughput Screening for Inducers of Myeloid Differentiation. *J Biomol Screen*. 2015;20(9):1150-1159.
94. Antony-Debre I, Paul A, Leite J, et al. Pharmacological inhibition of the transcription factor PU.1 in leukemia. *J Clin Invest*. 2017;127(12):4297-4313.
95. Morita K, Suzuki K, Maeda S, et al. Genetic regulation of the RUNX transcription factor family has antitumor effects. *J Clin Invest*. 2017;127(7):2815-2828.
96. Wang S, Kollipara RK, Srivastava N, et al. Ablation of the oncogenic transcription factor ERG by deubiquitinase inhibition in prostate cancer. *Proc Natl Acad Sci U S A*. 2014;111(11):4251-4256.

FIGURE LEGENDS

Figure 1 – Heptad regulatory regions have dynamic accessibility profiles across normal and leukemic blood development, and accessibility patterns are sufficient to classify normal and leukemic cells

(A) tSNE plot of single cell RNAseq in normal bone marrow, with cells labelled by inferred identity as determined by Setty et al 2019. HSC = haematopoietic stem cell, CLP = common lymphoid progenitor, DC = dendritic cell, Ery = erythoid lineage cells, Mega = megakaryocytes, Mono = monocyte lineage cells. (B) Relative expression of CD34 and heptad genes projected on to the tSNE plot in A. (C) Schematic of the branching hierarchy model of normal blood development showing relationships between the cell populations shown in D. (D) ATACseq peaks at heptad regulatory regions over developmental time. Plots show merged data from available replicates (Corces et al 2016). (E) ATACseq peaks at heptad regulatory regions in one representative AML patient, showing pre-leukemic HSCS (pHSC), leukemic stem cells (LCS), and leukemic blasts (Blast). (F) Classification of normal cell types using only ATACseq signal at heptad regulatory regions. Heatmap shows calculated distance between each sample and the training set. The red box indicates a single MEP replicate that was misclassified as a CMP. (G) Classification of AML nearest normal cell type using only ATACseq signal at heptad regulatory regions. Plots show distance from each normal cell type for pre-leukemic HSCS, LSCs, and leukemic blasts from seven AML patients. (H) Performance of the heptad regulatory region classifier compared to previous classification of these samples using genome wide enhancer cytometry (Corces et al 2016).

Figure 2 - A densely interconnected heptad autoregulatory circuit persists in AML cells with altered connectivity compared to CD34+ HSPCs.

(A) *Left:* ChIPseq binding pattern at heptad regulatory regions in $CD34^+$ HSPCs. Grey boxes indicate regulatory regions not computationally called as binding peaks for the indicated TF. Plots are scaled to 5x the height of the smallest called peak for that TF to allow visualisation of a wide range of peak heights. *Right:* Corresponding inferred heptad autoregulatory circuit. Most regulatory elements have all seven heptad TFs bound, * and bold border indicate regions where binding of a particular TF is absent. (B) *Left:* ChIPseq binding pattern at heptad regulatory regions in ME-1 AML cells. Grey boxes indicate regulatory regions not computationally called as binding peaks for the indicated TF. Plots are scaled to 5x the height of the smallest called peak for that TF to allow visualisation of a wide range of peak heights. *Right:* Corresponding inferred heptad autoregulatory circuit. Most regulatory elements have all seven heptad TFs bound, * and bold border indicate regions where binding of a particular TF is absent. (C) *Left:* ChIPseq binding pattern at heptad regulatory regions in KG-1 AML cells. Grey boxes indicate regulatory regions not computationally called as binding peaks for the indicated TF. Plots are scaled to 5x the height of the smallest called peak for that TF to allow visualisation of a wide range of peak heights. *Right:* Corresponding inferred heptad autoregulatory circuit. Most regulatory elements have all seven heptad TFs bound, * and bold border indicate regions where binding of a particular TF is absent

Figure 3 - Specific TF consensus binding motifs, particularly ETS and GATA motifs, are critical for function of heptad regulatory elements.

(A) Schematic showing process for selecting TF binding motifs for mutation, and luciferase reporter workflow. (B) *Left panel:* Schematics showing conserved TF binding motifs in heptad regulatory elements that were highly bound by heptad TFs in AML cell lines, and activity of wild type (WT) and mutated luciferase constructs in KG-1 and ME-1 cells.

Activity is scaled relative to the empty vector, and graphs show representative data from a single transfection experiment (* P < 0.05, ** P < 0.01, *** P < 0.001, t-test). *Right panel:* Heatmaps showing aggregate data from all luciferase experiments. Data from biological replicates were normalised to WT activity for each experiment, then aggregate data scaled relative to empty vector. Heatmaps are scaled from 0 to maximum luciferase activity for each regulatory element. (C) Schematics showing conserved TF binding motifs in heptad regulatory elements that were highly bound by heptad TFs in AML cell lines, and activity of WT luciferase constructs in KG-1 and ME-1 cells. Activity is scaled relative to the empty vector, and graphs show representative data from a single transfection experiment (* P < 0.05, ** P < 0.01, *** P < 0.001, t-test).

Figure 4. Single cell transcriptomics in ME-1 cells reveals branching heterogeneity consistent with GATA2 regulation.

(A) Cumulative expression distributions for heptad genes in single ME-1 cells. cppt: counts per ten thousand reads. (B) Pairwise Spearman correlations between heptad genes in single cells. (C) Censored distributions of gene expression for the gene pairs highlighted in B. The two lower panels show the expression of the second gene in the lowest 10% and highest 5% of expressing cells for the first gene. P values refer to a Kolmogorov-Smirnov 2-sample test between the purple and yellow distributions. (D) UMAP embedding of ME-1 cells and cell state assignment based on northstar (Zanini et al 2020) and the Palantir data as atlas (see Figure 1). Stream-lines show RNA velocity as computed by scVelo (Bergen et al 2019), projected onto the same embedding. *Inset:* Schematic of the branching phenotype within ME-1 cells, indicating the cell flux into the Ery-precursor-like state is a rare event. (E) Expression of four heptad genes highlighted in B on the embedding. Colour legend: purple = no

expression, green = low expression, yellow = high expression. (F) *Left:* Fold increase in heptad gene expression across the HSC to Ery-precursor-like state in ME-1 cells. *Right:* Fold increase in heptad gene expression across the HSC to Ery-precursor state in normal $CD34^+$ HSPCs cells. (G) Performance of random forest classifiers between HSC-like and Ery-precursor-like states in ME-1 trained solely on Palantir data with a spectrum of selected features. The presence of GATA2 expression in the model is essential for its accuracy. Error bars indicate SD over 10 runs of the predictor with data resampling in each run.

Figure 5 - Manipulating GATA2 and ERG in bulk ME-1 cells and normal $CD34^+$ HSPCs leads to altered heptad expression and can push cells towards the Ery-like state

(A) Schematic of the branching phenotype within ME-1 cells indicating relative expression of key heptad genes highlighted in Figure 4. (B) Effect of knocking down *GATA2*, *TAL1*, or *ERG* on heptad genes in ME-1 cells (error bars show 95% confidence interval). (C) *Left:* Effect of over-expressing GATA2 on heptad genes in ME-1 cells (RNAseq). *Right:* GSEA plots showing enrichment of genes associated with the Ery-precursor/Ery-precursor-like state in response to over-expressing GATA2 in ME-1 cells. (D) *Left:* Effect of knocking down *ERG* on heptad genes in CD34+ HSPCs (RNAseq). *Right:* GSEA plots showing enrichment of genes associated with the Ery-precursor/Ery-precursor-like state in response to knocking down *ERG* in $CD34^+$ HSPCs. FDR q-value for GSEA plots = 0 except where indicated by * q-value = 0.02. (E) Effect of knocking down *ERG* on heptad genes in $CD34^+$ HSPCs using two different shRNAS (error bars show 95% confidence interval). (F) *Left:* Colony forming capacity of $CD34^+$ cells transduced with control (shCON) or ERG (shERG, shERG-2) shRNAs. $CD34^+$ cells produce colonies derived from granulocyte and/or macrophage progenitor cells (CFU-GM; grey), multipotential progenitor cells (CFU-GEMM; black), and

erythroid progenitor cells (blast forming unit-erythroid (BFU-E); red). *Right:* Proportion of total colonies which are erythroid (BFU-E).

Figure 6 – Proposed model of heptad activity across haematopoietic differentiation.

(A) Heptad transcription factors form a densely interconnected network, with key regulatory elements accessible and heptad-bound in normal and leukemic stem cells. Accessibility of regulatory elements, and consequently heptad connectivity, is reduced as cells become more differentiated. (B) Schematics representing sc-RNAseq populations in normal and ME-1 cell populations. GATA2, TAL1, and ERG promote cell state changes along the HSC-Ery precursor axis in both normal CD34+ HSPCs and ME-1 cells.