ANALYSIS
https://doi.org/10.1038/s41587-019-0159-2

nature biotechnology

Corrected: Author Correction

A systems biology pipeline identifies regulatory networks for stem cell engineering

Melissa A. Kinney[1,2,3], Linda T. Vo[1,2,4], Jenna M. Frame[1,2,5], Jessica Barragan[1,2], Ashlee J. Conway[1,2], Shuai Li[6,7], Kwok-Kin Wong[6,7], James J. Collins[8,9,10,11], Patrick Cahan[12], Trista E. North[1,2,5], Douglas A. Lauffenburger[3*] and George Q. Daley[1,2,4,13*]

A major challenge for stem cell engineering is achieving a holistic understanding of the molecular networks and biological processes governing cell differentiation. To address this challenge, we describe a computational approach that combines gene expression analysis, previous knowledge from proteomic pathway informatics and cell signaling models to delineate key transitional states of differentiating cells at high resolution. Our network models connect sparse gene signatures with corresponding, yet disparate, biological processes to uncover molecular mechanisms governing cell fate transitions. This approach builds on our earlier CellNet and recent trajectory-defining algorithms, as illustrated by our analysis of hematopoietic specification along the erythroid lineage, which reveals a role for the EGF receptor family member, ErbB4, as an important mediator of blood development. We experimentally validate this prediction and perturb the pathway to improve erythroid maturation from human pluripotent stem cells. These results exploit an integrative systems perspective to identify new regulatory processes and nodes useful in cell engineering.

Stem cell biology, cell engineering and regenerative medicine often invoke developmental principles to differentiate cells toward target identities. However, much remains to be learned about how signaling pathways integrate to determine cell fate[1]. The past decade of cell engineering has shown that expression of individual genes, or sets of genes, is often insufficient to functionally reprogram cell identity[2,3], underscoring the need for new approaches to quantitatively describe and manipulate cell state. We previously established CellNet[4–6] to assess the fidelity of engineered cells by interrogating key gene regulatory networks (GRNs) that define native populations. CellNet extracts cell-type specific GRNs from transcriptional profiling data, compares the GRNs to those of bona fide primary cells and tissues to assign a similarity score, and identifies dysregulated transcriptional regulators that account for the differences between engineered cells and their native counterparts. The network-level CellNet algorithm confers robustness to biological and technical variability and encodes topological information about regulator-target relationships. A limitation of CellNet is that training data from a small number of terminal cell and tissue types obscure the phenotypic heterogeneity that arises during dynamic biological processes such as cell differentiation. More recent efforts have aimed to describe intermediate developmental states using trajectory-based methods, which employ cell–cell similarity metrics to infer dynamics[7–10]. However, these algorithms rely on single-cell transcriptomics to provide sufficiently powered datasets and largely forgo network analytics.

Here, we extend CellNet to quantitatively define network dynamics along a differentiation pathway. We show that publicly accessible gene expression datasets capture population-level differentiation states with high dynamic resolution and broad biological scope, including responses across a spectrum of experimental variables such as chemical and genetic perturbations. Our pipeline goes beyond the establishment of GRNs to enable quantification of differentiation dynamics and identification of key signaling pathways governing cell fate changes (Fig. 1a). We apply this general approach to characterize erythropoiesis, a dynamic process that generates red blood cells (RBCs) throughout the lifetime of the organism. We focused on this system because its temporal stages of differentiation, defined by distinct immunophenotypes, have been comprehensively characterized[11]. Our analyses confirm key processes involved in distinct stages of erythropoiesis and elucidate novel dynamic patterns of gene expression. To improve erythroid maturation in vitro, we constructed an interaction network connecting the dynamic molecular signatures that distinguish late erythroblasts from reticulocytes. Our network analytics identifies a role for ErbB signaling during erythropoiesis, which we validate in human, murine and zebrafish models and apply to the maturation of RBCs derived from human induced pluripotent stem cells (iPSCs).

Results

CellNet delineates stem cells and progeny. To analyze the dynamics of stem cell differentiation, we began by establishing GRNs for hematopoietic stem cells and differentiated progeny using CellNet as previously described[4–6]. We augmented the original CellNet compendium[4] of microarray datasets from 16 human cell and tissue types to include 164 publicly available erythroid microarray datas-

[1]Stem Cell Program, Boston Children's Hospital, Boston, MA, USA. [2]Division of Hematology/Oncology, Boston Children's Hospital and Dana Farber Cancer Institute, Boston, MA, USA. [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Harvard Medical School, Boston, MA, USA. [5]Department of Pathology, Beth Israel-Deaconess Medical Center, Boston, MA, USA. [6]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [7]Laura & Isaac Perlmutter Cancer Center, NYU Langone Medical Center, New York, NY, USA. [8]Institute for Medical Engineering & Science, Department of Biological Engineering, Synthetic Biology Center, Massachusetts Institute of Technology, Cambridge, MA, USA. [9]Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA, USA. [10]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [11]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA. [12]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [13]Harvard Stem Cell Institute, Boston, MA, USA. *e-mail: lauffen@mit.edu; George.Daley@childrens.harvard.edu

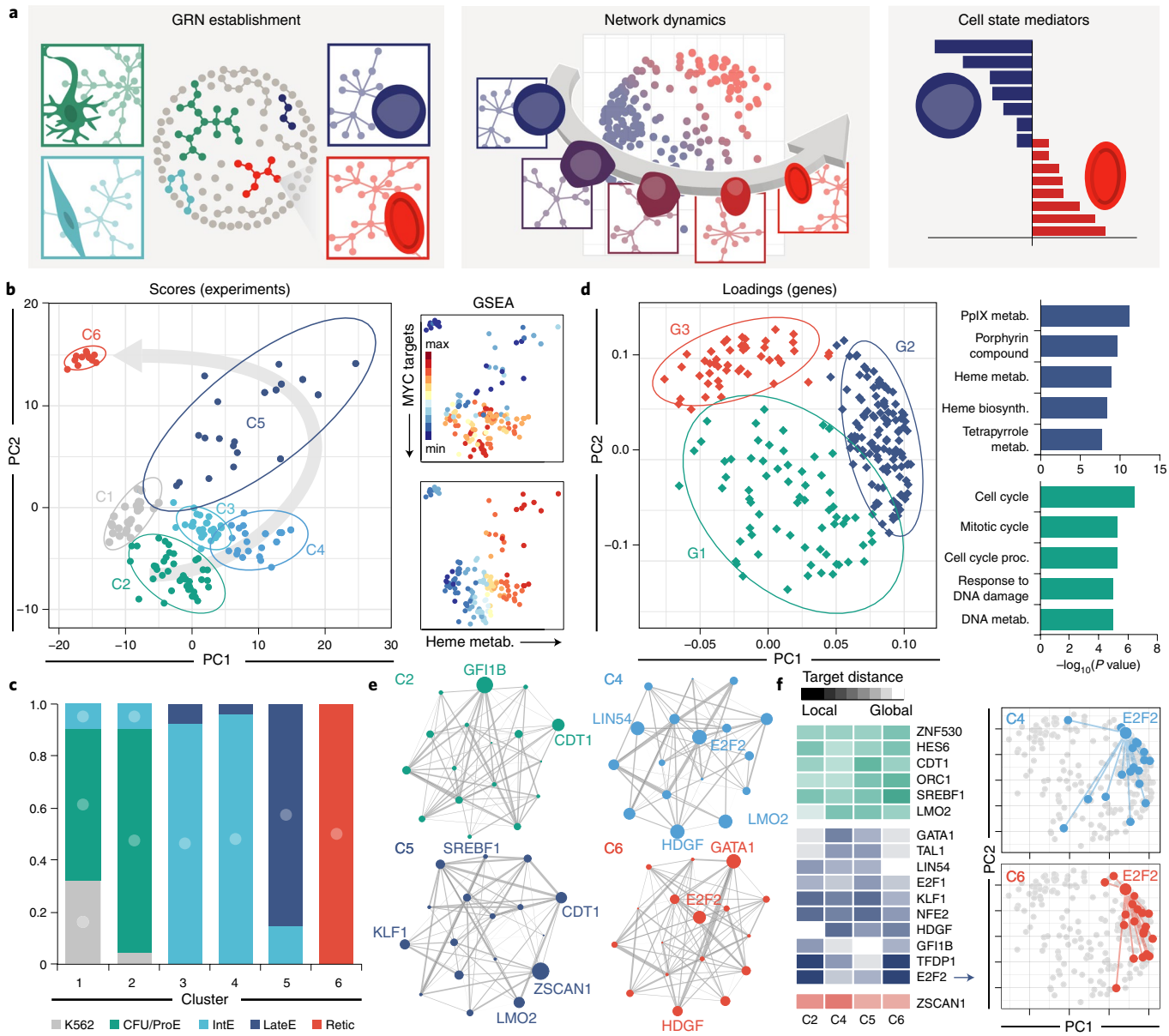

Fig. 1 | GRN dynamics capture cell fate specification. a, Schematic overview of computational pipeline. **b**, PCA scores of erythroid microarray datasets, with GMM derived clusters. **c**, Distribution of blinded manual classifications (K562 erythroleukemia cell line, early erythroblasts: CFU/ProE, intermediate erythroblasts: IntE, late/orthochromatic erythroblasts: LateE and reticulocytes) across unsupervised computational clusters C1–C6 from **b**, with circular annotations indicative of $P < 0.05$ by Fisher's exact test. **d**, PCA loadings plot for the 235 genes in the erythroid GRN and gene ontology biological process enrichment for GMM gene clusters, with G1–G3 correlating with early erythroblasts (green), intermediate/late erythroblasts (blue), reticulocytes (red), respectively. **e**, Visualization of regulators from dynamic networks calculated using CLR inference across each of the clusters from C2, C4, C5 and C6. Node size correlates with the degree (number of targets) and line width corresponds to the CLR *Z*-score (confidence of interaction) between regulators. **f**, The average Euclidean distance in the principal components loadings space was calculated for all targets of each regulator across the dynamic network topologies depicted in **e**, with representative plots demonstrating the target distribution for *E2F2* across networks from erythroblasts (C4) and reticulocytes (C6).

ets (Supplementary Tables 1 and 2). The erythroid data represented five manually classified phenotypes: early (CFU-E), intermediate (IntE) and late (LateE) erythroid progenitors, as well as reticulocytes and the K526 erythroleukemia cell line (Supplementary Table 3). Our rationale for augmenting the original compendium arose from the paucity and lack of biological variability in publicly available erythroid-specific sequencing datasets at the time of data compilation.

Application of the original CellNet classifier identified erythroid cells as hematopoietic stem and progenitor cells (HSPCs) with high probability (Supplementary Fig. 1a). However, after re-training the classifier with the augmented compendium and establishing an erythroid-specific GRN, CellNet robustly distinguished HSPCs and erythroid cell types (Supplementary Fig. 1b,c), with little overlap between the two GRNs other than cofactors mediating the canonical 'GATA switch'[12] that governs erythroid specification from HSPCs (Supplementary Fig. 1d,e,g). The erythroid GRN comprised 235 genes that were highly enriched for biological processes such as hemoglobin synthesis, oxygen transport, cell cycle and hematopoietic development (Supplementary Fig. 1f), with subnetworks

governed by 17 transcription factors, including canonical erythroid factors[13] such as *GATA1, LMO2, TAL1 and KLF1*. We also identified several factors not fully characterized in erythropoiesis, including *HES6* (ref. [14]), *CDT1* (ref. [15]) and *SREBF1* (ref. [16]). These data demonstrate that the CellNet algorithm can be readily augmented by the addition of new classifiers and accurately distinguishes biologically relevant GRNs, even among closely related cell types.

GRNs capture cell state dynamics. The augmented CellNet algorithm classified all stages of erythropoiesis with high probability (Supplementary Fig. 1h–j). We hypothesized that subnetworks, or smaller gene modules, within the erythroid GRN correspond to distinct stages of differentiation, analogous to the cell-type specific GRNs identified CellNet[4]. To dissect the erythroid GRN, we projected the data into a principal component space (Fig. 1b). PC1 (35.2% variance) was highly correlated with the gene set enrichment analysis (GSEA) Hallmark pathway for hemoglobin metabolism (Pearson's $r=0.91$; Fig. 1b and Supplementary Fig. 2a) and PC2 (19.5% variance) correlated with MYC targets (Pearson's $r=-0.8$; Fig. 1b and Supplementary Fig. 2b). Unsupervised Gaussian Mixture Model (GMM)–based clustering identified six discrete phenotypes, which were significantly enriched for manual, literature-based designations of erythroid stage—C1–C2: CFU-E/early proerythroblast; C3–C4: intermediate proerythroblast; C5: late erythroblast; C6: reticulocyte (Fig. 1c and Supplementary Table 3). K562, an erythroleukemia cell line, clustered in C1 and studies of hemoglobin-perturbed cells clustered in C3, a PC1-shifted intermediate cluster. We therefore focused on clusters C2, C4, C5 and C6 for the purpose of studying physiological and developmental erythropoiesis.

In contrast to whole genome-based dimensionality reduction techniques commonly used in trajectory algorithms, the erythroid GRN served as a feature selection upstream of PCA, which identifies genes correlated with developmental stages. By GMM clustering, erythroid network genes clustered into three distinct groups (Supplementary Table 4), with early (G1) and intermediate (G2) differentiation clusters associated with cell cycle and hemoglobin synthesis, respectively (Fig. 1d and Supplementary Table 5). The reticulocyte cluster (G3) comprised genes that were not significantly enriched for any biological processes. Likewise, ranking gene importance to each phenotypic cluster (Supplementary Fig. 2d–g and Supplementary Table 6) failed to yield annotations for the reticulocyte cluster (C6). We also implemented *K*-means clustering to identify sets of genes with similar dynamic expression across biological clusters (C2, C4, C5 and C6) and identified coordinated regulation of genes related to processes such as stress responses, autophagy and apoptosis during differentiation (Supplementary Fig. 2h).

We confirmed that dimensionality reduction similarly captured biologically meaningful clusters in developmentally staged, purified populations analyzed by bulk RNA-seq (GSE53983). Sample localization in the PCA space was driven by similar genes, as shown by significant correlations with biological analogs from microarray data—proerythroblast: C2/S1/S2; intermediate erythroblast: C4/S3; late erythroblast: C5/S4; positive Pearson's *r*, with $P<0.05$ (Supplementary Fig. 3c). Of note, expert knowledge was required for the interpretation and comparison of biologically analogous samples across microarray and RNA-seq datasets (for example, C4 and C5 both correlate with S3; C2 correlates with both S1 and S2), likely because the granularity and variance are strongly tied to the data source and experimental design. This further highlights the need for large data compendia, including purified populations, primary cells, in vitro differentiated cells, genetically perturbed cells and rare or unique populations (that is, reticulocytes), to fully sample the biological space within a given cell type. We also demonstrated that, in addition to compatibility with different data types, our pipeline is generally applicable to other biological systems (Supplementary Fig. 4 and Supplementary Tables 7–9).

To derive dynamic network models, we exploited topological regulatory information encoded in the erythroid GRN as a secondary layer atop the lineage-correlated loadings (Fig. 1e). Early erythropoiesis is dominated by a few, highly connected regulators, with more distributed regulation during the proerythroblast (C4) and late erythroblast (C5) stages (Fig. 1e). Canonical regulators, such as *GATA1* and *E2F2*, are also highly connected in the mature (C6) network, suggesting that early regulators impart persistent influence (Fig. 1f). *GATA1* has been implicated in erythroid maturation, with a distinct network from that of early erythroid specification[17]. *E2F2* also recurs as a central regulator in both the intermediate erythroblast (C4) and reticulocyte (C6) stages; however, there is a clear rewiring of its targets from a diffuse cluster of coregulated genes to a more compact network during maturation (Fig. 1f). Thus, the integration of network biology and GRN-based feature selection with dimensionality reduction uncovers dynamic changes in network activity and architecture accompanying cell fate changes.

Identification of pathways mediating cell fate transitions. We further explored the capacity for network analytics to identify biological processes that mediate stem cell differentiation. We focused on the late erythroblast (C5) to reticulocyte (C6) transition, as relatively little is known about the integrated mechanisms controlling terminal erythroid maturation. Moreover, microarray datasets derived from the in vivo reticulocyte transcriptome[18] (Supplementary Fig. 5) provided comparisons that are not readily accessible as these are transient, mobile populations.

To construct signatures of this transition, we employed the least absolute shrinkage and selection operator (LASSO) as a feature selection method that minimizes covariate correlation. The resulting 27-gene signature (Fig. 2a and Supplementary Table 10) accurately predicted the late erythroblast and reticulocyte cell states without overfitting, based on a partial least squares discriminant (PLSDA) model (Supplementary Fig. 6). This method produced a sparse gene set that lacked unifying annotations. We therefore adopted a 'bottom-up' approach using local network information to connect our signature genes (Fig. 2b). This propagation of LASSO targets is similar to network biology approaches to predict drug targets and disease-associated genes and is based on the hypothesis that genes in close proximity topologically are functionally related[19–21].

To identify common regulators, we investigated the local topology of the first-order subnetwork in the global CellNet GRN, from which cell-type specific GRNs were originally identified (Fig. 2c). Contrary to our hypothesis, there was largely a one-to-one connection between all connected regulators and LASSO targets (Fig. 2d). The few statistically enriched genes belonged to networks of coregulated transcription factors, such as the pluripotency factors (that is, *NANOG, SOX2, LIN28*)[22,23] that are associated with a single LASSO target, *SALL2*. This topology suggests that LASSO targets are associated with discrete biological processes, rather than being downstream of common regulators. Accordingly, this same analysis identified common regulators between ontologically related genes (Supplementary Fig. 7). However, further dissection of gene modules with modest coregulation of LASSO targets revealed that late erythroblast targets (Fig. 2e, Module 1) were associated with regulators of hematopoietic differentiation and P53-apopototic pathways, whereas reticulocyte LASSO genes (Fig. 2e, Module 2) were downstream of metabolic and lipid pathways important for RBC maintenance.

Based on this largely one-to-one topology of the transcriptional regulator-target network, we hypothesized that common signaling networks may lie upstream of the transcriptional layer. Therefore, we generated an interaction network using the STRING database (Fig. 2f, Supplementary Fig. 8 and Supplementary Table 11). We employed the Prize Collecting Steiner Forest (PCSF) algorithm[24], which is particularly suited for modeling multiple, independent pathways acting

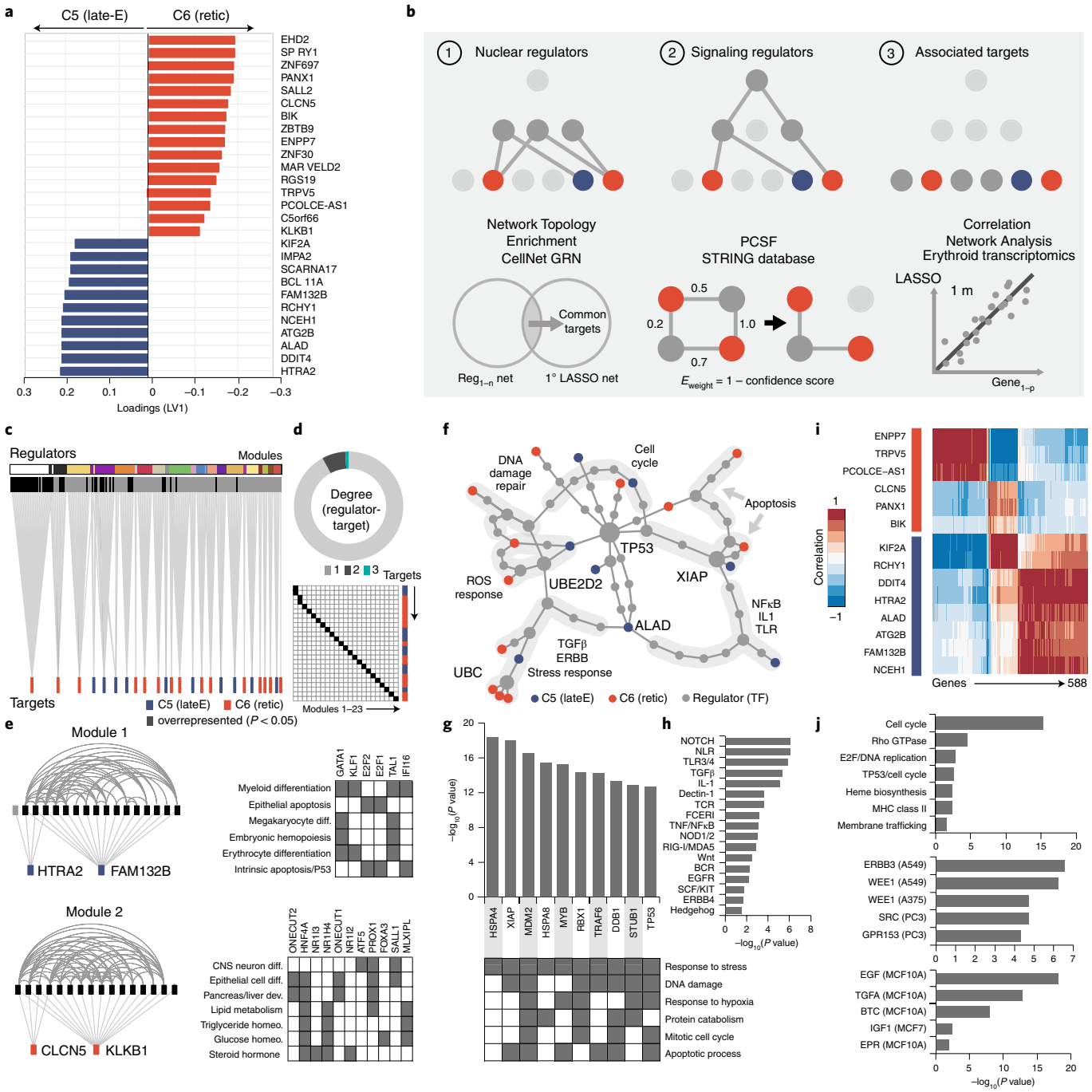

Fig. 2 | ErbB signaling is implicated in erythroid differentiation. a, Gene signature distinguishing clusters C5 (late-E; blue) and C6 (reticulocyte; red) from Fig. 2a, as determined by LASSO. **b**, schematic of network propagation approaches across different regulatory layers, including (1) nuclear regulators, (2), signaling regulators and (3) associated targets, corresponding to networks from **c**, **f** and **i**, respectively. **c**, First-order connection network derived from connecting the LASSO signature genes within the CellNet global GRN. Significantly enriched regulators ($P < 0.05$ by Fisher's test for connections in the network compared to the global GRN) are shown in black, with community-derived (that is, locally high density) modules depicted above in varying colors and the LASSO genes in red and blue, corresponding to the representation in **a**. **d**, The degree of all regulators and LASSO gene module membership. **e**, Graphical representation of Modules 1 and 2, with significantly enriched regulators (black) and the associated gene ontology biological process annotations. **f**, Protein–protein interaction network built from gene signature using the STRING database and the PCSF algorithm. Non-LASSO nodes (Steiner nodes) are depicted with the size proportional to the degree and LASSO nodes (terminals) are depicted in red and blue, corresponding to the representation in **a**. Modules within the network are highlighted and annotated with significantly enriched biological processes. **g**, *P* values ranking node enrichment (Fisher's test for connections in the LASSO network relative to the full STRING network) and corresponding gene ontology annotations. **h**, Enriched signaling pathways from the Reactome database. **i**, Coexpression network comprising genes highly correlated $r > |0.90|$ with the LASSO signature. **j**, Enrichment analyses for Reactome processes, kinase perturbation (downregulated genes on kinase knockdown from LINCS L1000) and ligand regulation (upregulated on ligand stimulation from LINCS L1000).

in synergy toward a unified biological response. The resulting network was enriched for biological processes such as apoptotic signaling, stress responses and cell cycle, consistent with previoius analyses (Fig. 2g). Uniting these processes, P53 is a highly interconnected central node ($P<0.001$, see Supplementary Fig. 8f). The network was also significantly enriched for Reactome signaling pathways relevant for erythropoiesis, such as Notch[25], Rho[26], TGFβ[27] and BCR[28], as well as novel candidate pathways, including epidermal growth factor receptor (EGFR)/ErbB4, TLR and RIG-I/MDA5 (Fig. 2h).

Finally, we used a 'guilt by association' approach to define networks that were highly correlated with the LASSO signature (Fig. 2i, Supplementary Fig. 9 and Supplementary Table 12). Highly enriched transcription factor binding (ENCODE and ChEA) and kinase regulation (LINCS L1000) further implicated proliferative and apoptotic processes (that is, E2F, P53 and FOXM1/WEE1). Moreover, several enriched kinases included members of the MAPK/ERK pathway (that is, SRC, ErbB3/ErbB4), and the ligand activation signatures (epidermal growth factor (EGF), TGFA, BTC) further supported a role for ErbB signaling in regulation of the coexpression network (Fig. 2j). This analysis demonstrates the utility of combining sparse gene signatures with network propagation approaches to identify novel biological processes that potentially mediate dynamic fate changes, thereby establishing hypotheses to be experimentally confirmed.

ErbB4 is necessary for efficient erythropoiesis. Although our network models identified several enriched signaling nodes and candidate pathways in erythroid maturation, the preponderance of evidence pointed to ErbB signaling. Although significantly enriched, EGFR/ErbB4 was not among the top candidate pathways (Fig. 2h, see comprehensive list in Supplementary Table 8); however, when combined with expert knowledge that ErbB signaling is frequently associated with P53 (ref. [29]), and the apoptotic[30] and proliferative[31] processes that were repeatedly identified in our network models, ErbB signaling emerged as a lead candidate. To determine whether ErbB signaling was necessary for erythroid maturation, we perturbed erythroblasts differentiated from bone marrow HSPCs ($CD34^+$) with ErbB inhibitors (Fig. 3a). Maturation ($GlyA^+CD71^-$) was only affected by pan-ErbB inhibitors (Afatinib, Dacomitinib, Neratinib), implicating ErbB4 rather than EGFR/ErbB2.

ErbB4 signaling has not previously been implicated in blood development or homeostasis. We characterized ErbB4 in human, mouse and zebrafish erythropoiesis. Using an erythroid in vitro differentiation protocol for human HSPCs[32], we observed increasing *ERBB4* mRNA expression as erythroid cells matured (Supplementary Fig. 10a). Native human bone marrow erythroid fractions also exhibited increased *ERBB4* expression in the most mature population ($GlyA^+CD71^-$) (Supplementary Fig. 10b). Reciprocally, pharmacological inhibition of ErbB signaling with Neratinib for 1 week shifted the bone marrow differentiation profile in mice, with an increase in immature and a decrease in mature erythroid populations (Fig. 3b), as well as changes to the peripheral hematopoietic fractions (Supplementary Fig. 11).

We next determined whether ErbB4 signaling also functioned during erythroid ontogeny, a process that initiates in multiple waves from restricted progenitors during embryogenesis[33]. Morpholino inhibition of ErbB4 in zebrafish embryos significantly decreased the frequency of $Gata1^+$ erythroid cells (Supplementary Fig. 12a) and of more differentiated globin-expressing cells (Fig. 3c,d) at 48–56 h post fertilization (hpf), without affecting neutrophils (Supplementary Fig. 12b). These data indicate that ErbB4 signaling is necessary for robust erythropoiesis during embryonic and adult hematopoiesis.

ErbB4 deficiency induces stress erythropoiesis. To more stringently characterize ErbB4 in adult erythropoiesis, we employed a genetic mouse model derived via αMHC-driven expression of human HER4 to circumvent embryonic lethality from heart defects in the whole body ErbB4 knockout ($ErbB4^{-/-}HER4^{heart}$)[34]. Consistent with the effects of Neratinib treatment, we observed an increase in early proerythroblast populations, with fewer mature orthochromatic and normoblastic cells in the $ErbB4^{-/-}$ bone marrow (Supplementary Fig. 13a,b). Nucleated RBCs and a high percentage of reticulocytes were present in peripheral blood, indicating moderate stress erythropoiesis in homozygotes and blood counts revealed significant changes in hemoglobin distribution (Fig. 4a and Supplementary Fig. 13c). $ErbB4^{-/-}$ mice had enlarged spleens (Supplementary Fig. 13d), a >two-fold expansion of early erythroblasts ($GlyA^+CD71^+$; gate II) (Fig. 4b) and overcrowded red pulp (Fig. 4c), suggesting extramedullary erythropoiesis. Morphological analysis demonstrated early developmental blocks across multiple lineages in $ErbB4^{-/-}$ bone marrow (Fig. 4d). $CD41^+$ megakaryocytes in ErbB4-deficient spleen ($P=0.007$ compared to wild type; Fig. 4e) decreased significantly, accompanied by a myeloid-skewed leukocyte profile and increased platelets in the periphery (Fig. 4f). These results demonstrate dysregulated multi-lineage hematopoietic phenotypes in $ErbB4^{-/-}HER4^{heart}$ mice.

Mitotic and proliferative processes downstream of ErbB matures iRBCs. To interrogate the molecular mechanisms downstream of ErbB, we performed global gene expression analysis of in vitro differentiated RBCs perturbed with pan versus selective inhibitors. Transcriptomic analysis confirmed that the erythroid GRN was modulated by pan-ErbB inhibitors, but not by Lapatinib, a dual EGFR/ErbB2 inhibitor (Supplementary Fig. 14a,b). Although cells were treated between the intermediate and late erythroblast stages, early network cluster genes (G1, see Supplementary Fig. 14c) were significantly decreased, suggesting that ErbB signaling plays a role during multiple stages of differentiation. Analysis of pathways dysregulated by pan-ErbB inhibition revealed upregulation of P53 signaling (Fig. 5a), with concomitant downregulation of mitotic and proliferative pathways (Fig. 5b). Consistent with our previous computational analysis, these data connect ErbB signaling with P53 and proliferative pathways in human erythropoiesis.

As mechanistic analyses identified the Wnt pathway as a putative downstream target of ErbB4 (Fig. 5b), we exploited the pharmacologic accessibility of this pathway to enhance erythropoiesis in vitro. A critical barrier to blood generation as a cell-based biotechnology stems from a block in erythroid maturation from iPS cells (iRBCs), often requiring the use of feeder cells. We promoted maturation of iRBCs in a feeder-free system using bioprocess-compatible hematopoietic progenitors, which undergo continuous expansion under doxycycline-induced overexpression of five transcription factors (Fig. 5c)[35]. Activation of Wnt signaling via the agonist CHIR99021 increased the maturation of iRBCs, resulting in a 1.8-fold ($P=4.8\times10^{-5}$) increase in $GlyA^+CD71^-$ orthochromatic erythroblasts (Fig. 5d). Concomitantly, cells decreased in size with an increased nuclear-to-cytoplasmic ratio (Fig. 5e). Collectively, these data demonstrate that systems-level identification of druggable signaling pathways in developmental processes, such as erythropoiesis, is directly applicable to stem cell biomanufacturing and regenerative cell therapies.

Discussion

Here, we establish the utility of systems-level analytics to elucidate biological processes that mediate dynamic stem cell and developmental transitions. Our computational pipeline provides a roadmap for the derivation of network models that connect sparse gene signatures with corresponding, yet disparate, biological processes, to captures the multi-factorial nature of cell state transitions. With cell engineering in hematopoiesis as an example, we highlight how to connect critical elements (for example, LASSO gene signatures)

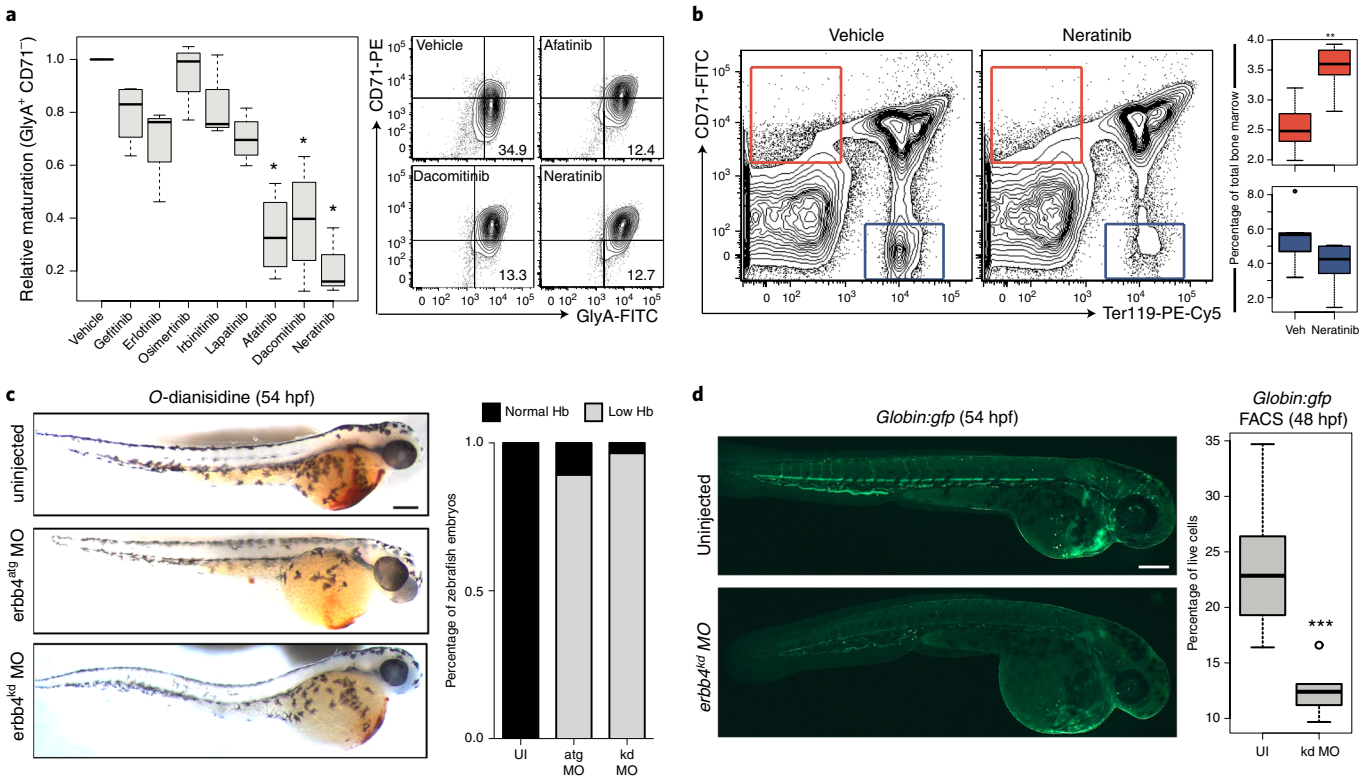

Fig. 3 | ErbB4 is required for robust erythroid development. a, Chemical screen with ErbB inhibitors during in vitro erythroid differentiation from bone marrow CD34$^+$ cells and quantified based on the relative fractions of GlyA$^+$CD71$^-$ cells, with representative flow cytometry plots. $^*P<0.05$ compared to vehicle (DMSO), via ANOVA and post hoc Tukey's HSD tests. $n=3$–4 independent experiments. **b**, Early (CD71$^+$GlyA$^-$) and late erythroid (CD71$^-$GlyA$^+$) progenitor profiles in the bone marrow of mice ($n=5$ per group) treated with vehicle (0.5% hydroxypropyl methylcellulose) or Neratinib (60 mg kg^{-1}) for 7 d. $^{**}P<0.01$ by two-sided, unpaired *t*-test. **c**, *o*-Dianisidine (3,3′-dimethoxybenzidine) staining of zebrafish 54 hpf after ErbB4 MO[51] injection (left). Quantification of the percentage of zebrafish embryos ($n>6$ across three clutches) with low hemoglobin (low *o*-dianidisine staining) relative to uninjected controls (right). **d**, Transgenic globin-GFP zebrafish embryos ($n>6$ across two clutches) were visualized and analyzed via flow cytometry at 54 and 48 hpf, respectively, after morpholino (MO) injection. $^{***}P<0.001$ by two-sided, unpaired *t*-test. Scale bars correspond to 200 µm.

to pathways/processes (for example, networks derived via PCSF and correlation). Our network models suggested and we experimentally confirmed a previously unanticipated role for ErbB4 in hematopoiesis.

Our advanced pipeline integrates network topological architecture with pseudotemporal information to provide multiple layers of information about cell differentiation, which is complementary to purely trajectory-based algorithms[7–10] and highlights the changing roles of transcriptional regulators across dynamic stages of development. Moreover, the LASSO feature reduction as a foundation for network modeling ensures that the resulting models are informed by genes most vital to distinguishing divergent cell states. In contrast to traditional differential gene expression approaches[19–21], LASSO produces a sparse, sharply focused gene set and, when combined with PCSF, produces a signaling network comprising branches associated with distinct biological processes. Together, these approaches provide a more global depiction of the systems-level processes associated with cell fate transitions.

By applying our pipeline to study hematopoietic specification, we established a role for ErbB4 signaling in erythropoiesis in multiple in vitro and in vivo models. Many of our computational approaches did not directly identify ErbB4; however, network propagation from our maturation signature repeatedly identified ErbB ligands and ErbB-associated signaling, including MAPK/ERK, mitotic processes, P53 and apoptosis[36,37]. This highlights the need for future development of unsupervised metrics to prioritize candidates from aggregate data, which currently requires expert knowledge as an integral part of the process. Although there were no annotated processes enriched within the reticulocyte gene cluster, it included the *N*-methyl-D-aspartate (NMDA) receptor, GRIN3B, which is commonly implicated, along with ErbB4, in neurological development[38] and pathophysiology[39]. Anemia is a common side effect of antipsychotic drugs[40] and studies of glutamate-mediated ion channels supports their functional role in erythropoiesis[41]. This opens the possibility of new avenues of crosstalk between neurological and hematopoietic systems, akin to the regulation of hematopoietic stem cell production by the central nervous system[42]. Our dynamic analyses also revealed that oxidative stress pathways peak at the late erythroblast stage; ErbB4 is a known stress responsive pathway in the heart[43] and abrogates oxidative damage in the brain[44]. Although a recent meta-analysis of genome-wide association study data identified neuregulin-4 (NRG4), an ErbB4-specific ligand, as a putative locus in aberrant human RBC phenotypes[45], the pathway has not been previously characterized in erythropoiesis.

Cell engineering has broadly focused on inducing transcription factors as the emissaries of phenotype. To this end, CellNet successfully predicts candidate and aberrant transcription factors. However, even the most-studied form of reprogramming, induced pluripotency, remains exquisitely sensitive to culture conditions and relies on signaling molecules, such as bFGF[46]. The advanced CellNet pipeline demonstrated here allows transcriptional targets to be complemented with druggable pathways. We demonstrate that the downstream ErbB signaling pathway can be exploited as a druggable target for more-robust production of RBCs from an iPSC-derived, bioprocess-compatible progenitor. Such multi-level reprogramming strategies may be especially beneficial for establishing and

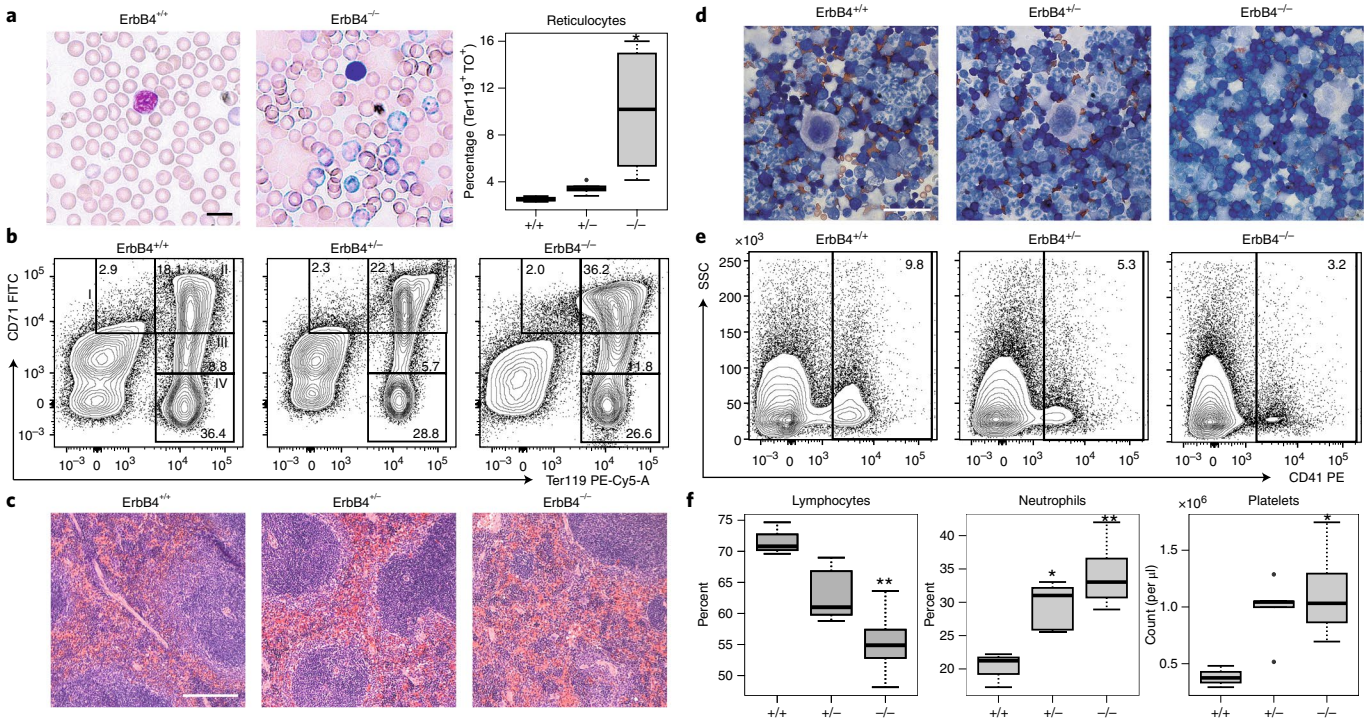

Fig. 4 | ErbB4 genetic deficiency leads to blood defects in the ErbB4$^{-/-}$HER4^{heart} mouse model. a, Morphology of the peripheral blood from smears stained with May–Grunwald–Giemsa indicating the increased reticulocyte fraction (purple), quantified as Ter119$^+$ Thiazole Orange (TO)$^+$ cells (ErbB4$^{+/+}$ $n=3$; ErbB4$^{-/-}$ $n=5$; ErbB4$^{+/-}$ $n=4$; $^*P<0.05$ compared to wild type by one-way ANOVA). **b**, Erythroid fractions within the spleen of wild type, heterozygous and homozygous mice, demonstrating increased extramedullary erythropoiesis (gate II, CD71$^+$Ter119$^+$). **c**, Morphology of the spleen stained with hematoxylin and eosin showing dense red pulp. **d**, Bone marrow composition from cytospins stained with May–Grunwald–Giemsa demonstrating the presence of immature fraction hematopoietic fractions across multiple lineages. **e**, Proportion of CD41$^+$ megakaryocytes in the spleen. **f**, Quantification of lymphocytes, neutrophils and platelets in the peripheral blood via Hemavet (ErbB4$^{+/+}$ $n=3$; ErbB4$^{-/-}$ $n=5$; ErbB4$^{+/-}$ $n=4$; $^*P<0.05$, $^{**}P<0.01$ compared to wild type by one-way ANOVA). Scale bars correspond to 10 μm (**a**), 100 μm (**c**) and 50 μm (**d**).

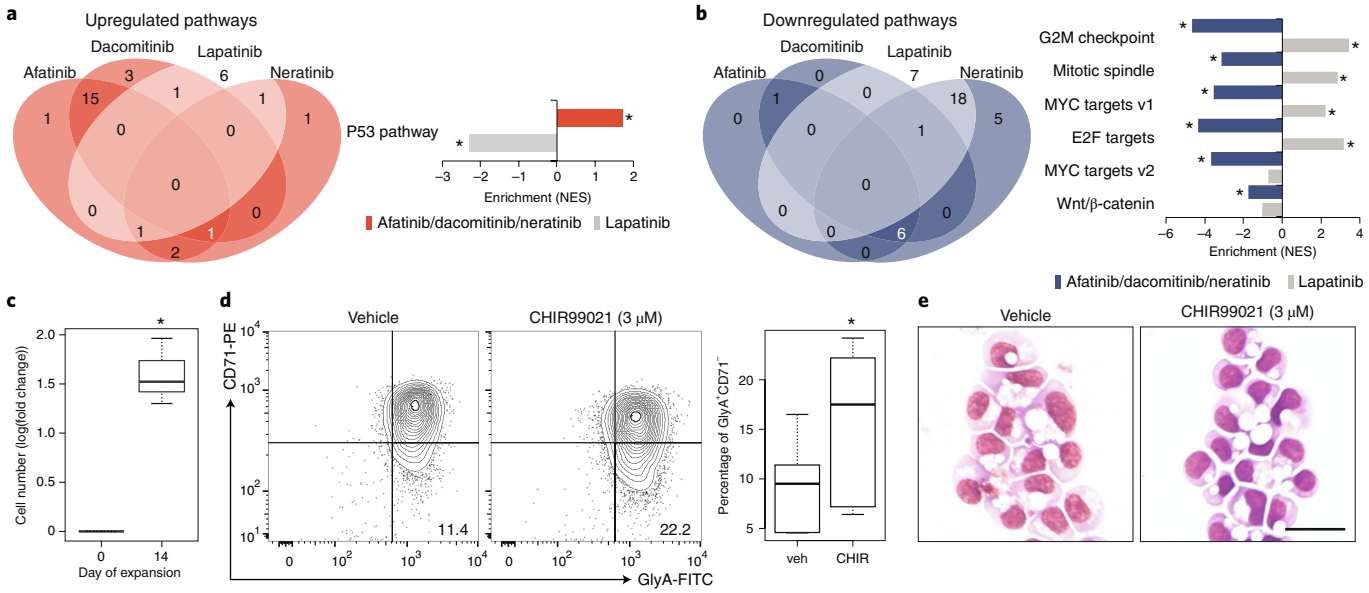

Fig. 5 | Modulation of pathways downstream of ErbB signaling augments iPSC-derived RBC generation. RNA-seq of bone marrow CD34$^+$ cells treated with Lapatinib (increased affinities for EGFR and HER2), as well as Neratinib, Dacomitinib and Afatinib (pan-ErbB inhibitors) for 24 h. **a**,**b**, Expression of genes from the Hallmark pathways (GSEA) significantly (*FDR < 0.25) upregulated (**a**) and downregulated (**b**) (compared to DMSO vehicle) and common to pan-ErbB inhibitor treated cells. **c**, Cell numbers (log fold change) of expanded iPS-derived CD34-5F cells[35] over 14 d. **d**, Erythroid profile with and without CHIR99021 treatment during the final week of differentiation, with significantly increased proportion of mature (GlyA$^+$, CD71$^-$) erythrocytes. $^*P<0.001$ by unpaired, two-sided *t*-test. **e**, morphology of May–Grunwald–Giemsa stained, iPS-derived orthochromatic erythroblasts. **c**, $n=7$ and **d**, $n=6$ independent replicates from two experiments. Scale bar in **e**, 20 μm.

maintaining elusive populations, such as HSCs. Although engraftable HSCs can be generated with transcription factors alone[47], reprogramming is enhanced by perturbation of developmental pathways, such as TGFβ and BMP4 (ref. [48]). Similarly, AKT-activated endothelial cells support self-renewal and maintenance of HSCs through angiocrine factors[49]. The prevalence of growth factor supplementation and cocultures across hematopoietic differentiation protocols further highlights the need to identify and recapitulate[50] cell-extrinsic signals.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0159-2.

Received: 21 December 2017; Accepted: 16 May 2019;
Published online: 1 July 2019

References

1. Westerhoff, H. V. & Palsson, B. O. The evolution of molecular biology into systems biology. *Nat. Biotechnol.* **22**, 1249–1252 (2004).
2. Morris, S. A. & Daley, G. Q. A blueprint for engineering cell fate: current technologies to reprogram cell identity. *Cell Res.* **23**, 33–48 (2013).
3. Hanna, J. H., Saha, K. & Jaenisch, R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. *Cell* **143**, 508–525 (2010).
4. Cahan, P. et al. CellNet: network biology applied to stem cell engineering. *Cell* **158**, 903–915 (2014).
5. Morris, S. A. et al. Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158**, 889–902 (2014).
6. Radley, A. H. et al. Assessment of engineered cells using CellNet and RNA-seq. *Nat. Protoc.* **12**, 1089–1102 (2017).
7. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
8. Shin, J. et al. Single-Cell RNA-Seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17**, 360–372 (2015).
9. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 1–11 (2014).
10. Lummertz da Rocha, E. et al. Reconstruction of complex single-cell trajectories using CellRouter. *Nat. Commun.* **9**, 892 (2018).
11. Dzierzak, E. & Philipsen, S. Erythropoiesis: development and differentiation. *Cold Spring Harb Perspect. Med.* **3**, a011601–a011601 (2013).
12. Tsai, F. Y. & Orkin, S. H. Transcription factor GATA-2 is required for proliferation/survival of early hematopoietic cells and mast cell formation, but not for erythroid and myeloid terminal differentiation. *Blood* **89**, 3636–3643 (1997).
13. Cantor, A. B. & Orkin, S. H. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene* **21**, 3368–3376 (2002).
14. da Cunha, A. F. et al. Global gene expression reveals a set of new genes involved in the modification of cells during erythroid differentiation. *Cell Prolif.* **43**, 297–309 (2010).
15. Ding, K. et al. Genetic Loci implicated in erythroid differentiation and cell cycle regulation are associated with red blood cell traits. *Mayo Clin. Proc.* **87**, 461–474 (2012).
16. Li, J. et al. Isolation and transcriptome analyses of human erythroid progenitors: BFU-E and CFU-E. *Blood* **124**, 3636–3645 (2014).
17. Rylski, M. et al. GATA-1-mediated proliferation arrest during erythroid maturation. *Mol. Cell Biol.* **23**, 5031–5042 (2003).
18. Goh, S.-H. et al. The human reticulocyte transcriptome. *Physiol. Genomics* **30**, 172–178 (2007).
19. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18**, S233–S240 (2002).
20. Langfelder, P., Mischel, P. S. & Horvath, S. When is hub gene selection better than standard meta-analysis? *PLoS ONE* **8**, e61505 (2013).
21. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
22. Takahashi, K. et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).
23. Yu, J. et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).
24. Tuncbag, N. et al. Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *J. Comput. Biol.* **20**, 124–136 (2013).
25. Robert-Moreno, A. et al. The notch pathway positively regulates programmed cell death during erythroid differentiation. *Leukemia* **21**, 1496–1503 (2007).
26. Watanabe, S. et al. Loss of a Rho-regulated actin nucleator, mDia2, impairs cytokinesis during mouse fetal erythropoiesis. *Cell Rep.* **5**, 926–932 (2013).
27. Tanno, T. et al. High levels of GDF15 in thalassemia suppress expression of the iron regulatory protein hepcidin. *Nat. Med.* **13**, 1096–1101 (2007).
28. Jacquel, A. et al. Apoptosis and erythroid differentiation triggered by Bcr-Abl inhibitors in CML cell lines are fully distinguishable processes that exhibit different sensitivity to caspase inhibition. *Oncogene* **26**, 2445–2458 (2007).
29. Zhou, B. P. et al. HER-2/neu induces p53 ubiquitination via Akt-mediated MDM2 phosphorylation. *Nat. Cell Biol.* **3**, 973–982 (2001).
30. Le, X. F. et al. Heregulin-induced apoptosis is mediated by down-regulation of Bcl-2 and activation of caspase-7 and is potentiated by impairment of protein kinase C alpha activity. *Oncogene* **20**, 8258–8269 (2001).
31. Holbro, T. et al. The ErbB2/ErbB3 heterodimer functions as an oncogenic unit: ErbB2 requires ErbB3 to drive breast tumor cell proliferation. *Proc. Natl Acad. Sci. USA* **100**, 8933–8938 (2003).
32. Lee, H.-Y. et al. PPAR-α and glucocorticoid receptor synergize to promote erythroid progenitor self-renewal. *Nature* **522**, 474–477 (2015).
33. Orkin, S. H. & Zon, L. I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell* **132**, 631–644 (2008).
34. Tidcombe, H. et al. Neural and mammary gland defects in ErbB4 knockout mice genetically rescued from embryonic lethality. *Proc. Natl Acad. Sci. USA* **100**, 8281–8286 (2003).
35. Doulatov, S. et al. Induction of multipotential hematopoietic progenitors from human pluripotent stem cells via respecification of lineage-restricted precursors. *Cell Stem Cell* **13**, 459–470 (2013).
36. Naresh, A. et al. The ERBB4/HER4 intracellular domain 4ICD is a BH3-only protein promoting apoptosis of breast cancer cells. *Cancer Res.* **66**, 6412–6420 (2006).
37. Bersell, K., Arab, S., Haring, B. & Kühn, B. Neuregulin1/ErbB4 signaling induces cardiomyocyte proliferation and repair of heart injury. *Cell* **138**, 257–270 (2009).
38. Li, B., Woo, R.-S., Mei, L. & Malinow, R. The neuregulin-1 receptor erbB4 controls glutamatergic synapse maturation and plasticity. *Neuron* **54**, 583–597 (2007).
39. Hahn, C.-G. et al. Altered neuregulin 1-erbB4 signaling contributes to NMDA receptor hypofunction in schizophrenia. *Nat. Med.* **12**, 824–828 (2006).
40. Flanagan, R. J. & Dunk, L. Haematological toxicity of drugs used in psychiatry. *Hum. Psychopharmacol.* **23**, 27–41 (2008).
41. Hänggi, P. et al. Functional plasticity of the N-methyl-D-aspartate receptor in differentiating human erythroid precursor cells. *Am. J. Physiol., Cell Physiol.* **308**, C993–C1007 (2015).
42. Kwan, W. et al. The central nervous system regulates embryonic HSPC production via stress-responsive glucocorticoid receptor signaling. *Cell Stem Cell* **19**, 370–382 (2016).
43. Kuramochi, Y. et al. Cardiac endothelial cells regulate reactive oxygen species-induced cardiomyocyte apoptosis through neuregulin-1beta/erbB4 signaling. *J. Biol. Chem.* **279**, 51141–51147 (2004).
44. Xu, Z. et al. Neuroprotection by neuregulin-1 following focal stroke is associated with the attenuation of ischemia-induced pro-inflammatory and stress gene expression. *Neurobiol. Dis.* **19**, 461–470 (2005).
45. van der Harst, P. et al. Seventy-five genetic loci influencing the human red blood cell. *Nature* **492**, 369–375 (2012).
46. Ludwig, T. E. et al. Derivation of human embryonic stem cells in defined conditions. *Nat. Biotechnol.* **24**, 185–187 (2006).
47. Sugimura, R. et al. Haematopoietic stem and progenitor cells from human pluripotent stem cells. *Nature* **545**, 432–438 (2017).
48. Lis, R. et al. Conversion of adult endothelium to immunocompetent haematopoietic stem cells. *Nature* **545**, 439–445 (2017).
49. Kobayashi, H. et al. Angiocrine factors from Akt-activated endothelial cells balance self-renewal and differentiation of haematopoietic stem cells. *Nat. Cell Biol.* **12**, 1046–1056 (2010).
50. Shukla, S. et al. Progenitor T-cell differentiation from hematopoietic stem cells using Delta-like-4 and VCAM-1. *Nat. Methods* **14**, 531–538 (2017).
51. Sato, T. et al. Neuregulin 1 type II-ErbB signaling promotes cell divisions generating neurons from neural progenitor cells in the developing zebrafish brain. *PLoS ONE* **10**, e0127360 (2015).

Acknowledgements

The authors thank G. Corfas (University of Michigan) for sharing the ErbB4$^{-/-}$ HER4heart mutant mice, which were generated in 2003 by M. Gassmann (University of Basel) and colleagues[34] and L.I. Zon (Boston Children's Hospital) for the *globin:eGFP* transgenic fish. The authors also thank P. Eser for the ErbB inhibitor library and T. Rosanwo for cells and reagents, as well as R. Mathieu and the BCH Flow Cytometry Core, J. Osborne, B. Joughin, J. Das and A. Zweemer for technical advice. This work is supported by grants to G.Q.D. from the NIH National Institute of Diabetes and Digestive and Kidney Diseases (nos. R24-DK092760, U54DK110805) and National Heart, Lung, and Blood Institute (Progenitor Cell Translation Consortium, no. U01HL134812; and

nos. R01-HL04880 and NIH R24-OD017870-01). Additional support was given to D.A.L. from NIH National Institute of General Medical Sciences (no. R01-GM069668). M.A.K. is supported by a NIH T32 Training Grant from BWH Hematology. L.T.V. was supported by the NSF Graduate Research Fellowship. J.M.F. is supported by a NIH T32 Training Grant from the NHLBI. T.E.N. is a Leukemia and Lymphoma Society Scholar. G.Q.D. is an associate member of the Broad Institute and was supported by the Howard Hughes Medical Institute and the Manton Center for Orphan Disease Research.

Author contributions

M.A.K., D.A.L. and G.Q.D. conceived the project. M.A.K., L.T.V., J.M.F., J.B., A.J.C. and S.L. performed experimental work and data interpretation. K.-K.W., J.J.C., P.C., T.E.N., D.A.L. and G.Q.D. supervised research and participated in project planning. M.A.K., T.E.N., D.A.L. and G.Q.D. prepared the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0159-2.

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

Correspondence and requests for materials should be addressed to D.A.L. or G.Q.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

GRN reconstruction and CellNet analytics. In total, 164 erythroid Affymetrix microarrays (Supplementary Tables 1 and 2) from the HGU133plus2 platform were acquired from the Gene Expression Omnibus (GEO) and compiled with the original human CellNet compendium[4]. Microarrays were preprocessed, the global GRN was calculated via the context likelihood of relatedness (CLR)[52] inference algorithm, and subnetworks were detected via InfoMap[53] community detection, as previously described[4,6]. Unless specified, all high dimensionality data analytics were accomplished using the R computational environment (v.3.2.2), with specified packages from CRAN and Bioconductor. All graphical representations and network analytics were visualized and calculated with the igraph package. Cell- and tissue-specific GRNs were established via enrichment using the chi-squared statistical test. As implemented in CellNet, a random forest classifier was trained based on the GRN for each cell type and trained with a randomly selected subset comprising approximately 50% of the microarrays. The classification performance was then evaluated on the remaining independent subset of microarrays. The sensitivity (true positive divided by the sum of true positives and false negatives) at a false positive rate of 5% was calculated as a metric to evaluate classifier accuracy. The GRN score, defined as the weighted mean of expression *Z*-score, was calculated as previously described[4].

PCA and trajectory establishment. To evaluate GRN dynamics, the 235 genes within the erythroid GRN were mean centered and scaled to unit variance before dimensionality reduction via decomposition of the multivariate dataset into principal component space via the *prcomp* command from the stats package. GSEA[54] enrichment scores were calculated for each sample as a pre-ranked list, relative to the population mean, and correlations with respect to the principal component axes were calculated via Pearson's *r*. All GSEA analyses were run using the Hallmark datasets in the Molecular Signatures Database (MSigDB)[55]. Unsupervised clustering of both microarrays and genes were calculated within the principal component space via GMM-based methods from the mclust package.

For global analysis of gene dynamics, differential genes were first identified via the following criteria: (1) expression above a minimum threshold of 3.5, (2) variation across the dataset using an interquartile range greater than 0.75 and (3) significance between clusters via Bonferroni adjusted analysis of variance (ANOVA). The remaining 1,788 differentially expressed genes were clustered via *K*-means into $K=12$ groups, with *K* determined by calculating Bayesian inference criteria. All enrichment analyses were conducted using standard chi-squared or Fisher statistics on gene ontology[56,57] defined biological processes.

LASSO and network propagation. The LASSO[58] was calculated within the glmnet package[59] using a binomial classification of microarrays from clusters C5 and C6 (Fig. 1b), based on the differential genes (1,788 genes) defined via criteria above. The value of lambda ($\lambda=0.009$), calculated via *cv.glmnet*, was chosen by minimizing the mean square error and corresponded to a signature of 27 genes. To validate the LASSO model, a two-component PLSDA model (mixOmics package[60]) was built based on the 27 genes and used to predict the binary classifications into clusters C5 and C6 (Supplementary Fig. 6). The calibration accuracy and error rates were calculated by the 'Leave One Out' method and compared to random models by: (1) shuffling the classifications and (2) selecting 27 random genes. In each case, the performance of random models was determined based on the average of 1,000 permutations.

To expand the network without overfitting, three additional models were built based on: (1) CellNet transcriptional regulatory networks, (2) protein–protein signaling networks and (3) coexpression networks (Fig. 2b). The CellNet first-order network was derived from the amalgamation of all first-order connections to the 27-gene signature within the global CellNet GRN. *P* values corresponding to the overrepresentation of each network regulator were calculated based on Fisher's test comparing the connections within the first-order network to those in the global GRN and corrected for multiple hypothesis testing. Regulatory modules within the first-order network were determined using the walktrap community algorithm.

The protein–protein interaction network was derived via querying the STRING database (v.9.0)[61,62] using the PCSF algorithm, as previously described[24]. Low confidence interactions with an edge score, $s(e)<0.5$ were removed and the cost was calculated as $1-s(e)$. For increased robustness, noise was added to the edge cost and the resulting network was the amalgamation of ten iterations, as previously described[63]. PCSF parameters, including μ (node degree penalty), ω (number of trees) and β (node prize scaling), were varied to demonstrate that the network generation was robust across a range of values (Supplementary Fig. 8). Larger networks (increasing ω and β), exhibited decreased density and centralization, with an increase in the number of significant ($P<0.05$) gene ontology annotations. This indicates that growing larger networks contributes toward the inclusion of discrete but cohesive biological processes, rather than adding random, unrelated genes/proteins. Moreover, smaller networks (Supplementary Fig. 8c–e) also exhibit common regulatory nodes consistent with our complete LASSO signaling network (Fig. 2f)

The coexpression network was determined by calculating the Pearson's *r* between each of the 27 signature genes and all other genes across clusters C2, C4, C5 and C6 (Fig. 1b). An absolute cutoff of 0.9 was selected for network reconstruction, based on the 'elbow' of network size over the range of thresholds (Supplementary Fig. 9a). The network parameters, as well as gene ontology annotations were calculated across the full range of cutoff thresholds and representative networks spanning the range exhibit similar features in terms of predicted transcription factors (ChEA/ENCODE), kinases (LINCS L1000) and ligands (LINCS L1000) (Supplementary Fig. 9b–e). Enrichment analyses for the all networks were queried via gene ontology[56,57] and Enrichr[64] in their native implementations.

Statistical analyses. All statistical analyses were calculated in R, using two-sided, unpaired *t*-test, ANOVA or Fisher's exact test. Data are presented as standard boxplots representing the median and ranging from the 25th to 75th percentiles, with the whiskers extending to 1.5× interquartile range. The sample sizes represent a minimum of three independent replicates, corresponding to distinct experiments and/or parallel biological replicates (for example, animals or cell cultures). The exact replicate numbers and statistical tests are specified in the figure legends.

Human CD34$^+$ RBC differentiation. Human CD34$^+$ progenitors derived from mobilized peripheral blood (AllCells) were expanded for 4 d in StemSpan SFEM (StemCell Technologies) with the addition of interleukin (IL)-3 (10 ng ml^{-1}), IL-6 (50 ng ml^{-1}), thrombopoietin (TPO) (50 ng ml^{-1}), stem cell factor (SCF) (50 ng ml^{-1}) and Flt3 (50 ng ml^{-1}). Unless specified, all cytokines were from PeproTech. Erythroid differentiation was accomplished using a previously published, three-stage protocol[32]. Briefly, all stages of differentiation consisted of a basal erythroid differentiation medium (EDM) consisting of: IMDM with 15% FBS, 1% BSA, 2 mM L-glutamine, 500 µg ml^{-1} holo-transferrin and 10 µg ml^{-1} insulin. Stage 1 consists of EDM plus the addition of dexamethasone (1 µM), β-estradiol (1 µM), IL-3 (5 ng ml^{-1}), SCF (100 ng ml^{-1}) and erythropoietin (EPO) (6 U) for 5 d (days 0–5). Stage 2 consists of EDM plus the addition of SCF (50 ng ml^{-1}) and EPO (6 U) for 4 d (days 5–9). Stage 3 consists of EDM plus the addition of EPO alone (2 U) for 8 d (days 9–17). At all stages, cells are cultured in 24-well plates in 1 ml of media. Cell number seeded at the beginning of stages 1, 2 and 3 are: 10^5, 2×10^5 and 5×10^5 per well.

iPS-5F generation and RBC differentiation. Human iPS-5F cells were generated as previously described[35,65] from MSC-iPS[66] obtained from the Boston Children's Hospital Human Embryonic Stem Cell Core and verified by immunohistochemistry for pluripotency markers, teratoma formation and karyotyping. Briefly, iPS cells were differentiated as embryoid bodies using a hematopoietic induction protocol[67] and CD34$^+$ cells were sorted from bulk embryoid body culture by magnetic activated cell sorting using human CD34 microbeads (Miltenyi Biotec.), as per the manufacturer's instructions. The embryoid body progenitors were seeded on retronectin-coated (10 µg cm^{-2}) 96-well plates (2×10^4–5×10^4 cells per well) in SFEM (StemCell Technologies) containing 50 ng ml^{-1} SCF, 50 ng ml^{-1} FLT3, 50 ng ml^{-1} TPO (all R&D Systems), 50 ng ml^{-1} IL-6 and 10 ng ml^{-1} IL-3 (both from PeproTech) and infected with 5F lentiviral particles. Lentiviral particles for the 5F plasmids (HOXA9, ERG, RORA, SOX4 and MYB cloned into pInducer-21 doxycycline-inducible vector) were produced by transfecting 293T-17 cells (ATCC) with third-generation packaging plasmids. The multiplicity of infection (MOI) for each factor was: ERG MOI = 5, HOXA9 MOI = 5, RORA MOI = 3, SOX4 MOI = 3, MYB MOI = 3. Following 24 h of infection, 5F cells were cultured in SFEM with 50 ng ml^{-1} SCF, 50 ng ml^{-1} FLT3, 50 ng ml^{-1} TPO, 50 ng ml^{-1} (all R&D Systems) IL-6 and 10 ng ml^{-1} IL-3 (PeproTech) and 2 µg ml^{-1} doxycycline (Dox, Sigma). Cultures were maintained at a density of $<1\times10^6$ cells ml^{-1} and the medium was changed every 3–4 d.

RBC differentiation from iPS-5F followed a slightly modified protocol that was previously optimized for translational approaches aimed at transfusion of in vitro-derived RBCs[68]. In this protocol, the EDM was instead consisting of: IMDM with 5% inactivated plasma (solvent detergent pooled plasma AB from the Rhode Island Blood Center), 2 mM L-glutamine, 330 µg ml^{-1} holo-transferrin and 10 µg ml^{-1} insulin, 2 IU per ml heparin (Sigma) and 3 IU per ml EPO. Stage I (days 0–7) was plated at $1–3\times10^5$ cells per ml and supplemented with 10 µM hydrocortisone, 100 ng ml^{-1} SCF and 5 ng ml^{-1} IL-3. Stage II (days 7–11) was plated at $1–3\times10^5$ cells per ml and supplemented with 100 ng ml^{-1} SCF. Stage III (days 11–18) was plated at 1×10^6 cells per ml in the basal EDM. All analyses were conducted at day 18 of differentiation and CHIR99021 (3 µM) was added throughout stage III (days 9 and 13).

Flow cytometry and cell sorting. Human erythropoiesis, including differentiation from bone marrow CD34$^+$ cells and native bone marrow samples, was analyzed with the following antibody panel: CD71 PE (M-A712, BD) and CD235a/Glycophorin A PE- Cy7 (11E4B-7-6; Coulter) or CD235a/Glycophorin A FITC (11E4B-7-6, Coulter). Mouse erythropoiesis from Neratinib treated and HER4^{heart} mice was analyzed with the following antibody panel: mCD71 FITC (C2, BD), mTer119 PE-Cy5 (Ter119; eBioscience). All staining was performed with $<1\times10^6$ cells per 100 µl staining buffer (PBS + 2% FBS) with 1:100 dilution of each antibody for 30 min at room temperature in the dark. Compensation was performed by automated compensation with anti-mouse Igk and negative beads

(BD). Acquisition was performed on a BD Fortessa cytometer and all sorting was performed on a BD FACS Aria II cell sorter using a 70-mm nozzle. Gating strategies are depicted in Supplementary Fig. 15.

Inhibitors. All inhibitors were added to cell cultures at 1 μM on days 9 and 13 of differentiation, corresponding to the beginning and middle of stage 3 (supplemented with EPO only, as described above). DMSO was used for a vehicle control in all cell culture studies. Details on ordering information and affinities are provided in Supplementary Table 13.

RNA-sequencing. RNA was extracted after 24 h of incubation with ErbB inhibitors (day 10 of erythroid differentiation) using Trizol reagent (Invitrogen) and the RNeasy Plus kit (Qiagen). Quality of RNA was monitored via QC for high RNA integrity number values and low levels of DNA contamination. RNA-seq libraries were prepared using the SMARTseq v.4 kit as per the manufacturer's protocol with 10 ng input RNA. Libraries were sequenced using the 200-cycle paired-end kit on the Illumina HiSeq2500 system. RNA-seq reads were analyzed with the Tuxedo Tools following a standard protocol on the Harvard Medical School Orchestra Cluster. Reads were mapped with TopHat v.2.1.0 and Bowtie2 v.2.2.4 with default parameters against build hg19 of the human genome, and build hg19 of the RefSeq human genome annotation. Samples were quantified with the Cufflinks package v.2.2.1. Differential expression was performed using Cuffdiff with default parameters.

PCR. RNA was extracted as described above and complementary DNA was synthesized using the SuperScript VILO cDNA Synthesis Kit (Thermo), per the manufacturer's instructions. Real-time PCR was run using SYBR green technolgoy with QuantiTect primers for the ErbB receptor family (Qiagen) on the QuantStudio Flex Real-Time PCR System.

Zebrafish studies. Zebrafish were maintained according to institutional animal care and use committee-approved protocols. The *Tg(globin:eGFP)* line was provided by L.I. Zon, Children's Hospital, Harvard Medical School, Boston. MOs (GeneTools) were microinjected at the one-cell stage as described previously[69]. ErbB4 MOs were generated from previously published sequences[51]. Embryos were harvested at 48–56 hpf and were processed with matched sibling controls for *o*-dianisidine staining and evaluation of globin:eGFP intensity. Staining intensity was categorized as low, medium or high in that experiment, as previously described[70]; effects were independently confirmed by other laboratory members.

Mouse studies. All mice were housed in pathogen-free animal facilities, and all experiments were performed with the approval of the Animal Care and Use Committee at Harvard Medical School and the Dana Farber Cancer Institute and/or the BCH Animal Care Committee. At least $n=3$ animals were used per cohort, based on previous studies. For drug treatments, mice were assigned randomly to groups and not blinded. Neratinib was delivered to B6 albino mice via oral gavage at 60 mg kg^{-1} daily for 1 week. Hydroxypropyl methylcellulose was used as a vehicle control for Neratinib in mouse treatments.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All RNA-seq data have been deposited to the GEO database under GSE108128.

References

52. Faith, J. J. et al. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**, e8 (2007).
53. Rosvall, M. & Bergstrom, C. T. An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl Acad. Sci. USA* **104**, 7327–7331 (2007).
54. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
55. Liberzon, A. et al. The molecular signatures database hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).
56. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
57. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
58. Tibishirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
59. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
60. Lê Cao, K.-A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253 (2011).
61. Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28**, 3442–3444 (2000).
62. Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362–D368 (2017).
63. Akhmedov, M. et al. PCSF: an R-package for network-based interpretation of high-throughput data. *PLoS Comput. Biol.* **13**, e1005694 (2017).
64. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
65. Vo, L. T. et al. Regulation of embryonic haematopoietic multipotency by EZH1. *Nature* **553**, 506–510 (2018).
66. Park, I.-H. et al. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141–146 (2008).
67. Chadwick, K. et al. Cytokines and BMP-4 promote hematopoietic differentiation of human embryonic stem cells. *Blood* **102**, 906–915 (2003).
68. Giarratana, M.-C. et al. Proof of principle for transfusion of in vitro-generated red blood cells. *Blood* **118**, 5071–5079 (2011).
69. North, T. E. et al. Prostaglandin E2 regulates vertebrate haematopoietic stem cell homeostasis. *Nature* **447**, 1007–1011 (2007).
70. Doulatov, S. et al. Drug discovery for Diamond-blackfan anemia using reprogrammed hematopoietic progenitors. *Sci. Transl. Med.* **9**, eaah5645 (2017).

nature research

Corresponding author(s): George Q. Daley

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	164 erythroid Affymetrix microarrays (Supplementary Tables 1&2) from the HGU133plus2 platform were acquired from the Gene Expression Omnibus (GEO) and compiled with the original human CellNet compendium. Microarrays were preprocessed, the global gene regulatory network (GRN) was calculated via the Context Likelihood of Relatedness (CLR) inference algorithm, and subnetworks were detected via InfoMap community detection, as previously described. Unless specified, all high dimensionality data analytics were accomplished using the R computational environment (version 3.2.2), with specified packages (see Online Methods) from CRAN and Bioconductor.

Data analysis	All statistical analyses were calculated in R, using two-sided, unpaired t-test or ANOVA, where appropriate, with $p<0.05$ considered significant. Data are presented as standard boxplots representing the median and ranging from the 25th to 75th percentiles, with the whiskers extending to 1.5*IQR. The sample sizes represent a minimum of three independent replicates, as specified in the figure legends. Gene set enrichment analysis (GSEA) was run according to default parameters in their native implementations. Statistical enrichment of gene lists in gene ontology (GO) enrichment was performed using Fisher's exact test followed by correction for multiple hypothesis testing. RNA-seq reads were analyzed with the Tuxedo Tools following a standard protocol on the Harvard Medical School Orchestra Cluster. Reads were mapped with TopHat version 2.1.0 and Bowtie2 version 2.2.4 with default parameters against build hg19 of the human genome, and build hg19 of the RefSeq human genome annotation. Samples were quantified with the Cufflinks package version 2.2.1. Differential expression was performed using Cuffdiff with default parameters.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-seq data has been deposited to the Gene Expression Omnibus (GEO) database under GSE108128. Raw data is available in the Supplementary Tables.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	At least n=5 mice were used per cohort for neratinib treatments. For zebrafish studies, at least n>6 embryos were used and replicated across multiple clutches. HER4heart mouse studies were analyzed with at least n=4.
Data exclusions	No data were excluded
Replication	Number of replicates per experiment and statistical analyses are described in each figure legend.
Randomization	Mice and zebrafish were assigned randomly to groups and not blinded.
Blinding	Mice and zebrafish were assigned randomly to groups and not blinded.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Human erythropoiesis, including differentiation from BM CD34+ cells and native bone marrow samples, was analyzed with the following antibody panel: CD71 PE (M-A712; BD), and CD235a/Glycophorin A PE- Cy7 (11E4B-7-6; Coulter) or CD235a/ Glycophorin A FITC (11E4B-7-6; Coulter). Mouse erythropoiesis from Neratinib treated and HER4heart mice was analyzed with the following antibody panel: mCD71 FITC (C2; BD), mTer119 PE-Cy5 (Ter-119; eBioscience).
Validation	Antibodies were all previously validated using cord blood and peripheral blood mononuclear cells as positive controls.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	iPS-5F cells were generated as previously described (Doulatov et al., Cell Stem Cell, 2014; Vo et al., Nature, 2017) from MSC-iPS (Park et al., Nature, 2008) obtained from the Boston Children's Hospital Human Embryonic Stem Cell Core (hESC) and verified by immunohistochemistry for pluripotency markers, teratoma formation and karyotyping. Primary cells (CD34+ cord blood) were obtained from AllCells.
Authentication	MSC-iPS were verified by immunohistochemistry for pluripotency markers, teratoma formation and karyotyping
Mycoplasma contamination	All lines routinely tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	N/A

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	All mice were housed in pathogen-free animal facilities, and all experiments were performed with the approval of the Animal Care and Use Committee at Harvard Medical School and Dana-Farber Cancer Institute and/or the BCH animal care committee. At least n=4 animals were used per cohort. Mice were assigned randomly to groups and not blinded. Neratinib was delivered to B6 albino mice via oral gavage at 60 mg/kg daily for 1 week. Hydroxypropyl methylcellulose (HPMC) was used as a vehicle control for Neratinib in mouse treatments. ErbB4-/-HER4heart mutant mice were generated in 2003 by Martin Gassmann (University of Basel) and colleagues and obtained from Dr. Gabriel Corfas at the University of Michigan.
Wild animals	N/A
Field-collected samples	N/A

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	CD34+ progenitors derived from mobilized peripheral blood (AllCells) were expanded and differentiated via an established erythroid differentiation protocol. All staining was performed with < 1x106 cells per 100 µL staining buffer (PBS + 2% FBS) with 1:100 dilution of each antibody for 30 min at RT in dark. Compensation was performed by automated compensation with anti-mouse Igk and negative beads (BD).
Instrument	Acquisition was performed on a BD Fortessa cytometer and all sorting was performed on a BD FACS Aria II cell sorter using a 70-mm nozzle.
Software	All flow cytometry data was analyzed using FlowJo 8.7.
Cell population abundance	No sorting was performed.

Gating strategy

All cells were first gated on FSC/SSC according to cell size and granularity, using stained human cord blood mononuclear cells (MNCs) as a positive control and reference for cell size, granularity and staining intensity. Unstained samples were used to set up negative gates. Dead cell populations were excluded using DAPI staining.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.