Supplementary text

Estimating the discrete probability distribution of a motif variable

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

For example, three Ebox motifs were found at Erg+65 (Fig. 3a). They can be bound by either Scl or Lyl1. Thus, we assigned that *P*(Ebox@Erg+65=1) and *P*(Ebox@Erg+65=2) were determined by {3, Scl, Lyl1}. We assumed that (i) the expression level of a TF is proportional to the probability of that TF binding to a target motif; and (ii) the bindings of TFs to multiple motifs are independent events. Gene expression levels were defined within the closed interval [0, 1], which is identical to the possible range of probabilities. For ease of calculation, we took the expression level of a TF as its probability of binding to a motif. Accordingly, we have

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

$$\begin{aligned}\tilde{P}(\text{Ebox@Erg}+65=2)=&\textstyle\sum_{n=1}^{3}C(3,n)\times p^n\times(1-p)^{(3-n)}\times(1-q)^3\\&+\textstyle\sum_{n=1}^{3}C(3,n)\times q^n\times(1-q)^{(3-n)}\times(1-p)^3\\&+\textstyle\sum_{n=1}^{2}\sum_{m=1}^{3-n}C(3,n)\times p^n\times(1-p)^{(3-n)}\times C((3-n),m)\times q^m\times(1-q)^{(3-m)}\end{aligned} \tag{2}$$

where p and q represent the expression levels of Scl and Lyl1, respectively.

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

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

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

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

It should be mentioned that the number of the same motifs in a regulatory region was directly taken into account in the estimation of probabilities. One may raise the question of whether this number has such strong power. Specifically, should the exponents in equations (1) and (2) change linearly, or less than linearly, along with the increase in the number of Ebox motifs? To address this issue, we replaced all exponents with their square roots and rerun the whole set of simulations (data not shown). Results showed that using the square roots instead of the original numbers (i) caused a more evenly distributed expression of the nine TFs over the hypothetical interval [0, 1], (ii) captured the same trend in gene expression changes in some perturbations (e.g. the AML-ETO simulation), but (iii) led to decreased expression levels of certain TFs in other perturbations (e.g. PU.1 knockdown and Gfi1b over-expression), which therefore disagrees with the experimental data. In order to capture a better agreement of computational and experimental results, we directly used the number of motifs to estimate the discrete probability distributions.

Estimating the activity of a regulatory region

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

Based on this information, we estimated the regression coefficient of the Erg+65 region on a relevant motif family in the following way:

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

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

$$\tilde{y} = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 \tag{7}$$

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

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

$$\tilde{y} = c + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \varepsilon \tag{8}$$

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

Estimating the expression level of a gene

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

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

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

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

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

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

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

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

References from Supplementary Data

1. Beck D, Thoms JA, Perera D, Schütte J, Unnikrishnan A, Knezevic K, et al. Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. Blood. 2013;122(14):e12-22.

2. Bee T, Ashley EL, Bickley SR, Jarratt A, Li PS, Sloane-Stanley J, et al. The mouse Runx1 +23 hematopoietic stem cell enhancer confers hematopoietic specificity to both Runx1 promoters. Blood. 2009;113(21):5121-4.

3. Bee T, Swiers G, Muroi S, Pozner A, Nottingham W, Santos AC, et al. Nonredundant roles for Runx1 alternative promoters reflect their activity at discrete stages of developmental hematopoiesis. Blood. 2010;115(15):3042-50.

4. Chan WY, Follows GA, Lacaud G, Pimanda JE, Landry JR, Kinston S, et al. The paralogous hematopoietic regulators Lyl1 and Scl are coregulated by Ets and GATA factors, but Lyl1 cannot rescue the early Scl-/- phenotype. Blood. 2007;109(5):1908-16.

5. Göttgens B, Broccardo C, Sanchez MJ, Deveaux S, Murphy G, Göthert JR, et al. The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5' bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. Mol Cell Biol. 2004;24(5):1870-83.

6. Göttgens B, Ferreira R, Sanchez MJ, Ishibashi S, Li J, Spensberger D, et al. cis-Regulatory remodeling of the SCL locus during vertebrate evolution. Mol Cell Biol. 2010;30(24):5741-51.

7. Göttgens B, Nastos A, Kinston S, Piltz S, Delabesse EC, Stanley M, et al. Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. EMBO J. 2002;21(12):3039-50.

8. Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat Cell Biol. 2013;15(4):363-72.

9. Nottingham WT, Jarratt A, Burgess M, Speck CL, Cheng JF, Prabhakar S, et al. Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. Blood. 2007;110(13):4188-97.

10. Pimanda JE, Ottersbach K, Knezevic K, Kinston S, Chan WY, Wilson NK, et al. Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. Proc Natl Acad Sci U S A. 2007;104(45):17692-7.

11. Sánchez M, Göttgens B, Sinclair AM, Stanley M, Begley CG, Hunter S, et al. An SCL 3' enhancer targets developing endothelium together with embryonic and adult haematopoietic progenitors. Development. 1999;126(17):3891-904.

12. Sinclair AM, Göttgens B, Barton LM, Stanley ML, Pardanaud L, Klaine M, et al. Distinct 5' SCL enhancers direct transcription to developing brain, spinal cord, and endothelium: neural expression is mediated by GATA factor binding sites. Dev Biol. 1999;209(1):128-42.

13. Swiers G, Baumann C, O'Rourke J, Giannoulatou E, Taylor S, Joshi A, et al. Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level. Nat Commun. 2013;4:2924.

14. Wilkinson AC, Kawata VK, Schütte J, Gao X, Antoniou S, Baumann C, et al. Single-cell analyses of regulatory network perturbations using enhancer-targeting TALEs suggest novel roles for PU.1 during haematopoietic specification. Development. 2014;141(20):4018-30.

15. Wilson NK, Miranda-Saavedra D, Kinston S, Bonadies N, Foster SD, Calero-Nieto F, et al. The transcriptional program controlled by the stem cell leukemia gene Scl/Tal1 during early embryonic hematopoietic development. Blood. 2009;113(22):5456-65.

16. Wozniak RJ, Boyer ME, Grass JA, Lee Y, Bresnick EH. Context-dependent GATA factor function: combinatorial requirements for transcriptional control in hematopoietic and endothelial cells. J Biol Chem. 2007;282(19):14665-74.