Materials and Methods

Animals

All experimental protocols were reviewed and approved by the Institutional Animal Care and Use Committee of Tokyo Women's Medical University (approval ID: 13-99-2-B). Mice were purchased from Sankyo Labo Service.

Cell collection

$CD34^-$ KSL ($c\text{-}kit^+Sca1^+Lin^-$) LT-HSCs or $CD34^+$ KSL ST-HSCs were sorted, as described previously [36]. In brief, we isolated bone marrow cells from 8- to 10-week-old C57BL/6 mice and stained them with antibodies for CD34 (RAM34, eBiosciences, San Diego, CA), Sca-1 (E13-161.7, BD Biosciences Pharmingen, San Jose, CA), c-kit (2B8, BD Biosciences Pharmingen), and a lineage marker (Lineage Detection Kit, Miltenyi Biotec Inc., Bergisch Gladbach, Germany). Subsequently, we analyzed the stained cells using a MoFlo XDP cell sorter system (Beckman Coulter, Fullerton, CA).

RNA sequencing and real-time PCR

After obtaining total RNA extracts from 5000 LT- or ST-HSCs using Isogen (Nippon Gene, Tokyo, Japan) in triplicate, we synthesized cDNA using a SMARTer Pico cDNA amplification kit (Clonetech, Mountain View, CA) and amplified them with 20 cycles of PCR. Using the standard protocols for the SOLiD system, we sequenced the amplified cDNA using a SOLiD sequencer (Life Technologies, Carlsbad, CA), as described previously [36]. In the RT-PCR assay, total RNA was obtained from the sorted cells and cDNA was synthesized as described above. We performed RT-PCR using a TaqMan Gene Expression Assay (Life Technologies) for the genes indicated with the BioMark HD system (Fludigm, South San Francisco, CA).

Read mapping and quantification

We used the TopHat (v1.4.1)/Cufflinks (v.2.0.2) pipeline [33] with the sequenced reads (quality score, >15). The pipeline was coupled to Bowtie (v.0.12.7) [62]. We employed the recursive read mapping method, as described previously [32]. In brief, we applied TopHat by truncating the 3′ ends of unmapped reads and by realigning the reads using more stringent parameters. We set the parameters empirically, which were used sequentially, as the read length, "-initial-read-mismatches", "-segment-mismatches", and "-segment-length": (50, 3, 2, 25), (46, 3, 2, 23), (42, 3, 2, 21), (38, 2, 0, 19), and (34, 2, 0, 17).

The pipeline, which quantifies RNA abundance as fragments per kilobase of exon per million mapped reads (FPKM), mapped sequenced reads to the mouse genome (mm9), and then assembled transcripts with uniquely mapped reads (uni-reads) for each replicate. We used Cuffcompare to merge all the transcript assemblies; 14,728 and 14,128 RefSeq-annotated genes in LT- and ST-HSCs, respectively. Using the merged transcript assembly, we performed Cuffdiff, which calculates FPKMs across all replicates and detects DEGs via two-group *t*-tests coupled to a Benjamini–Hochberg false discovery rate (FDR) procedure. We further used transcripts that satisfied the following conditions: successful deconvolution, FDR of <0.05, complete match of intron chain, and FPKM of >0.001. The mouse genome and RefSeq annotation were downloaded from http://genome.ucsc.edu/.

Long-term competitive reconstitution assay

We cultured $CD34^-$ KSL HSCs derived from C57BL/6-Ly5.1 congenic mice for 5 days with or without 20μM GW1929 (Sigma-Aldrich, St. Louis, MO) in S-Clone SF-03 medium (Sanko-Junyaku Co., Tokyo, Japan) supplemented with 0.5% bovine serum albumin (Sigma, St. Louis, MO) and 50 ng/ml mouse stem cell factor and 50 ng/ml mouse TPO (all from R&D systems, Minneapolis, MN). Next, we performed a long-term competitive reconstitution assay by transplanting cultured cells with 5×10^5 whole bone marrow competitor cells derived from C57BL/6-Ly5.2 Wt mice into lethally irradiated (9.5 Gy) C57BL/6-Ly5.2 Wt mice.

Log-linear model (LLM)

Suppose that we consider binary-stated (absence or presence) TFs $\{A, B, C\}$. The observed counts fall into 2^3-dimensional contingency table by cross-classifying the TF states. The full model (FM), which contains all the possible interactions, gives the logarithms of probabilities as follows:

$$\log p_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}, \quad (1)$$

where i, j and k are the state indices of $\{A, B, C\}$, λs are unknown parameters, λ_{ij}^{AB}, λ_{ik}^{AC} and λ_{jk}^{BC} represent the interaction effects among the indexed variables. If an instance of A is independent of B, FM can be reduced to a reduced model (RM) with respect to the hierarchy [31], which is given as follows:

$$\log p_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ik}^{AC} + \lambda_{jk}^{BC}. \quad (2)$$

This model can be reformulated as

$$p_{ijk} = (p_{i+k} \cdot p_{+jk})/p_{++k}, \quad (3)$$

where "+" denotes the summation over the corresponding index. This formula is equivalent to $\Pr(A=i,B=j|C=k) = \Pr(A=i|C=k.)\Pr(B=j|C=k)$, which means that A and B are independent in the conditional distribution given C ($A \perp\!\!\!\perp B|C$).

To find the most parsimonious RM, we remove an interaction term from the current model and measure two p-values for the asymptotic χ^2 test of a likelihood ratio G^2 statistic [31]. The p-values comprise p_FM, which is the difference between FM and RM, and p_RM, which is the difference between the current model and RM. We accept a removal if it yields the largest p_RM (≥ 0.01), and we terminate if any removal test yields <0.01 for either p_RM or p_FM.

Iterative random sampling for LLM

A large number of TFs can easily yield a vast dimensional contingency table. To find a near optimal parsimonious model even in such higher-dimensional space, we designed an iterative sampling scheme that allowed us to calculate interaction probability Pr as follows.

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is an undirected graph, where $\mathcal{V}$ is a finite set of vertices (TFs) and $\mathcal{E}$ is a set of edges, which represent the interactions between vertex pairs. The scheme is as follows.

1. 1. $\mathcal{S} = \{s_1, \ldots, s_k\}$, a nonredundant combination of TFs, is selected randomly from all TFs ($k=10$ in the present study).
2. For all possible vertex pairs (s_i, s_j), the trial number $ntry_{ij}$ of an edge between s_i and s_j is counted (i.e., FM of k variables).

3. LLM infers the best model $\mathcal{G}' = (\mathcal{S}, \mathcal{E}')$, where $\mathcal{E}'$ is a set of edges that represents TF–TF interactions.
4. For all possible vertex pairs (s_i, s_j), if an edge in $\mathcal{E}'$ links a pair (s_i, s_j), the observed edge frequency $nobs_{ij}$ for this pair is counted.
5. For all possible vertex pairs (s_i, s_j), the interaction probability Pr for a pair (s_i, s_j) is updated using $nobs_{ij}/ntry_{ij}$.
6. If $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E}$ is a set of edges ($Pr = 1.0$), is not changed with a large number of samplings ($=100{,}000$); therefore, this procedure is terminated. Otherwise, steps 1–5 are repeated.

Linear regression model

We used a multivariate regression model

$$\log Y_i = \sum_j w_j X_{ij} + e_i, \tag{4}$$

$$X_{ij} = \sum_k x_k, \tag{5}$$

where Y_i is the expression of gene i, X_{ij} is TGAS of the jth TFBS in the promoter region of gene i, w_j is RC of the jth TFBS, and e_i is the error term. TGAS is the sum of scores x_k, where k represents the position of the jth TFBS in promoter i. We tested the following forms of x_k.

- I: matrix similarity s of TFBS j scored using MATCH[43] ($x_k = s_k$).
- II: TGAS I modified by a location-dependent weight L,

$$x_k = s_k \times L_k. \tag{6}$$

- III: TGAS II weighted by the expression fold change (F) of TFs,

$$x_k = s_k \times L_k \times \sum_{k'} F_{k'}, \tag{7}$$

where k' is the index of TFs binding to TFBS j. If FPKM for TF is ≤ 3, we use $F = 1$.

- IV: the same as TGAS III, but we removed TFBSs where none of the TFs had FPKM of >3.
- V: TGAS III weighted using both Fs of interactive TFs and the interaction probability Pr estimated by LLM,

$$x_k = s_k \times L_k \times \left(\sum_{k'} F_{k'} + I_{k'}\right) \tag{8}$$

$$I_{k'} = \sum_{i'=1}^{k'} \sum_{j'>i'}^{k'} F_{i'} F_{j'} Pr_{i'j'}. \tag{9}$$

We used a published method to calculate L [40]. First, we calculated the distribution of TFBS j in bins ($=500$ bp) of promoter regions and created a histogram H_{real}. Next, we randomized the positions of TFBS j and created a histogram H_{rand}. L for the kth TFBS j is given by the following:

$$L_k = \begin{cases} 0, & \text{if } H_{real}(m) < H_{rand}(m) \\ \dfrac{H_{real}(m) - H_{rand}(m)}{H_{real}(m)}, & \text{if } H_{real}(m) \geq H_{rand}(m), \end{cases} \tag{10}$$

where m represents the index of bin that corresponds to the position of the kth TFBS j. This location-dependent weight takes a value between 0 and 1, where a higher weight implies nonrandom occurrence.

Stepwise selection of the regression model

We built a regression model with the explanatory variable X and then reduced the model using AIC. Let the reduced model be Y' with X'. $X - X' = \{x_1, x_2, \ldots\}$ is the variables removed on the basis of AIC. V is the set of all pairwise terms of $x_i x_j$ ($i \neq j$). We searched any elements of V that improve Pearson's correlation coefficient r of 5-fold CV on testing datasets.

1. Randomly select v_i ($\in V$) and add it to X', which yields X''.
2. Perform 5-fold CV with X'' and calculate the averaged r on testing datasets.
3. If the r has been improved, update X'' to X'.
4. Repeat step 1–3 until all v_i have been tested.
5. Calculate Pearson's correlation coefficient R between observed and predicted FPKMs of all genes by using the final model.

We run this procedure 100 times using different random seeds. The final R is referred to as a model quality in this study.

Bioinformatics analysis

We obtained array-based gene expression profiles [8,9] from BloodExpress [63], RNA-seq data for megakaryocyte/erythroid precursors and megakaryocytes from http://genome.ucsc.edu/encode/, and RNA-seq data for MII oocytes and two-cell embryos from DDBJ DRA001066. The public RNA-seq datasets were analyzed using the pipeline mentioned above. To search putative TFBSs and TFs in TRANSFAC professional (released in January 2013) [39], we prepared ± 5kb DNA sequences from transcription start sites (TSSs) annotated in RefSeq (http://www.ncbi.nlm.nih.gov/refseq/), and applied the MATCH tool in the minimize false-positive mode [43].

To analyze the enriched GO terms, we used the DAVID Bioinformatics Resources [35]. Significant terms detected by DAVID (EASE score, a modified Fisher's exact p-value, <0.01) were grouped into representative ancestor terms in the dataset GO Slim2 using CateGOrizer [64]. We used the R programming language (http://www.r-project.org/) for regression modeling and to perform statistical tests. Although all p-values were adjusted by Bonferroni correction (Tables S6 and S8–S11), we used uncorrected p-values throughout this study to avoid too conservative interpretation that would reduce biologically meaningful findings.

Data access

The RNA-seq data generated in this study have been deposited in the DDBJ (DNA Data Bank of Japan) Sequence Read Archive (DRA) under accession number DRA001213. The online version of LLM is available at http://dbtmee.hgc.jp/tools/.

Supporting Information

Figure S1 Correlation analysis of gene expression levels measured using RNA-seq assays. (A) Reproducibility based on triplicate analyses of LT- and ST-HSCs. (B) Comparison of the gene expression correlations in the present study to those reported by Karlsson et al. [15], who purified HSCs using $CD48^-$, $CD150^+$, $CD34^-$, $CD9^{high}$ KSL for LT-HSCs and $CD48^-$, $CD150^+$, $CD9^{low}$ KSL for ST-HSCs.
(EPS)

Figure S2 Contribution of higher-order TF interaction scores estimated by LLM. (A) Statistical differences of 2 regression coefficient (RC) ensembles of a TFBS found commonly by TGAS III and V (two-sample *t*-test). (B) Distribution of the TF interaction score I_k in Equation 9.
(EPS)

Figure S3 Box plots of RCs estimated by 100 iterations of regression modeling with TGAS V. Pos and Neg represent the positive (red) and negative (blue) mean values of RCs (red line), respectively.
(EPS)

Figure S4 Subnetworks involved in ST-HSC regulation. Although the majority of TF-coding genes found in ST-HSCs (Figure 4A) were not differentially expressed, 26 differentially expressed TFs that putatively bind to 21 TFBSs were present among DEGs (Class A and Class B).
(EPS)

Figure S5 Propensity of the TFBS activities inferred from public RNA-seq datasets. We applied our method to public RNA-seq datasets related to sequential cell development (A) and lineage commitment (C). Our procedure evaluates the averaged R of 5-fold CV on testing datasets (blue line). If a model improved R in testing, the model was accepted and its R value between the observed and predicted gene expression of all genes was measured (red line). (B) Of 147 TFBSs ($p<0.05$), 67 TFBSs (Class A; upregulated in Oo) and 80 TFBSs (Class B; upregulated in 2C) exhibited significant gains and losses of activity ($p<0.001$). In addition, 73% (49/67) of Class A and 52.5% (42/80) of Class B genes exhibited no changes in the effects of their TFBS activities between cells, i.e., positive (negative) in Oo was still positive (negative) in 2C. We found that 16% (8/49) of Class A and 83% (35/42) of Class B genes had increased activities in 2C compared with Oo. (D) Among 150 TFBSs ($p<0.05$), 98 TFBSs (Class A, upregulated in MEP) and 114 TFBSs (Class B, upregulated in Mk) exhibited significant gains and losses of activity ($p<0.001$). We also found that 83% (81/98) of Class A and 76% (87/114) of Class B genes exhibited no changes in the effects of their TFBS activities. All of the TFBSs in both classes exhibited increases in the strengths of their activities in Mk compared with MEP. R, Pearson's correlation coefficient; Oo, MII oocytes; 2C, 2-cell embryo; MEP, megakaryocyte/erythroid precursor; Mk, megakaryocyte.
(EPS)

Table S1 RNA-seq mapping statistics.
(XLSX)

Table S2 Differentially expressed cell-surface molecules.
(XLSX)

Table S3 Differentially expressed transcription factors.
(XLSX)

Table S4 Transcription factors categorized into Class C.
(XLSX)

Table S5 Low expressed transcription factors (Class D).
(XLSX)

Table S6 Average regression coefficient of 142 TFBSs.
(XLSX)

Table S7 Classification of MkE, GM, and Lymphoid-associated genes.
(XLSX)

Table S8 TFBSs significantly different in the regression coefficient between LT- and ST-HSCs (Class A).
(XLSX)

Table S9 TFBSs significantly different in the regression coefficient between LT- and ST-HSCs (Class B).
(XLSX)

Table S10 Enriched GO terms in Class A.
(XLSX)

Table S11 Enriched GO terms in Class B.
(XLSX)

Table S12 Result of log-linear model in Class A.
(XLSX)

Table S13 Result of log-linear model in Class B.
(XLSX)

Acknowledgments

We thank Drs S. Mitani and T. Furukawa for their great help with RNA sequencing. Computational resources were provided by the supercomputer system at Human Genome Center, the Institute of Medical Science, the University of Tokyo.

Author Contributions

Conceived and designed the experiments: SJP KN. Performed the experiments: TU YS MY. Analyzed the data: SJP MSA. Contributed reagents/materials/analysis tools: TU SJP. Wrote the paper: SJP.

References

1. Hoang T (2004) The origin of hematopoietic cell type diversity. Oncogene 23: 7188–98.
2. Forsberg EC, Bhattacharya D, Weissman IL (2006) Hematopoietic stem cells: expression profiling and beyond. Stem Cell Rev 2: 23–30.
3. Sanjuan-Pla A, Macaulay IC, Jensen CT, Woll PS, Luis TC, et al. (2013) Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. Nature 502: 232–6.
4. Yamamoto R, Morita Y, Ooehara J, Hamanaka S, Onodera M, et al. (2013) Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. Cell 154: 1112–26.
5. Passegue E, Wagers AJ, Giuriato S, Anderson WC, Weissman IL (2005) Global analysis of proliferation and cell cycle gene expression in the regulation of hematopoietic stem and progenitor cell fates. J Exp Med 202: 1599–611.
6. Forsberg EC, Prohaska SS, Katzman S, Heffner GC, Stuart JM, et al. (2005) Differential expression of novel potential regulators in hematopoietic stem cells. PLoS Genet 1: e28.
7. Zhong JF, Zhao Y, Sutton S, Su A, Zhan Y, et al. (2005) Gene expression profile of murine longterm reconstituting vs. short-term reconstituting hematopoietic stem cells. Proc Natl Acad Sci U S A 102: 2448–53.
8. Mansson R, Hultquist A, Luc S, Yang L, Anderson K, et al. (2007) Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors. Immunity 26: 407–19.
9. Ficara F, Murphy MJ, Lin M, Cleary ML (2008) Pbx1 regulates self-renewal of long-term hematopoietic stem cells by maintaining their quiescence. Cell Stem Cell 2: 484–96.

10. Kent DG, Copley MR, Benz C, Wohrer S, Dykstra BJ, et al. (2009) Prospective isolation and molecular characterization of hematopoietic stem cells with durable self-renewal potential. Blood 113: 6342–50.
11. Chotinantakul K, Leeanansaksiri W (2012) Hematopoietic stem cell development, niches, and signaling pathways. Bone Marrow Res 2012: 270425.
12. Kunisaki Y, Bruns I, Scheiermann C, Ahmed J, Pinho S, et al. (2013) Arteriolar niches maintain haematopoietic stem cell quiescence. Nature 502: 637–43.
13. Liu P, Barb J, Woodhouse K, Taylor JG, Munson PJ, et al. (2011) Transcriptome profiling and sequencing of differentiated human hematopoietic stem cells reveal lineage-specific expression and alternative splicing of genes. Physiol Genomics 43: 1117–34.
14. Lu R, Neff NF, Quake SR, Weissman IL (2011) Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat Biotechnol 29: 928–33.
15. Karlsson G, Rorby E, Pina C, Soneji S, Reckzeh K, et al. (2013) The tetraspanin cd9 affords high-purity capture of all murine hematopoietic stem cells. Cell Rep 4: 642–8.
16. Weishaupt H, Sigvardsson M, Attema JL (2010) Epigenetic chromatin states uniquely define the developmental plasticity of murine hematopoietic stem cells. Blood 115: 247–56.
17. Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, et al. (2010) Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell 7: 532–44.
18. Bissels U, Bosio A, Wagner W (2012) Micrornas are shaping the hematopoietic landscape. Haematologica 97: 160–7.
19. Whichard ZL, Sarkar CA, Kimmel M, Corey SJ (2010) Hematopoiesis and its disorders: a systems biology approach. Blood 115: 2339–47.
20. Bonzanni N, Garg A, Feenstra KA, Schutte J, Kinston S, et al. (2013) Hard-wired heterogeneity in blood stem cells revealed using a dynamic regulatory network model. Bioinformatics 29: i80–8.
21. Hannah R, Joshi A, Wilson NK, Kinston S, Gottgens B (2011) A compendium of genome-wide hematopoietic transcription factor maps supports the identification of gene regulatory control mechanisms. Exp Hematol 39: 531–41.
22. Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, et al. (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell 144: 296–309.
23. Will B, Vogler TO, Bartholdy B, Garrett-Bakelman F, Mayer J, et al. (2013) Satb1 regulates the self-renewal of hematopoietic stem cells by promoting quiescence and repressing differentiation commitment. Nat Immunol 14: 437–45.
24. Mirshekar-Syahkal B, Haak E, Kimber GM, van Leusden K, Harvey K, et al. (2013) Dlk1 is a negative regulator of emerging hematopoietic stem and progenitor cells. Haematologica 98: 163–71.
25. Gazit R, Garrison B, Rao T, Shay T, Costello J, et al. (2013) Transcriptome analysis identifies regulators of hematopoietic stem and progenitor cells. Stem Cell Reports 1: 266–280.
26. Osawa M, Hanada K, Hamada H,·Nakauchi H (1996) Long-term lymphohematopoietic reconstitution by a single cd34-low/negative hematopoietic stem cell. Science 273: 242–5.
27. Ema H, Morita Y, Yamazaki S, Matsubara A, Seita J, et al. (2006) Adult mouse hematopoietic stem cells: purification and single-cell assays. Nat Protoc 1: 2979–87.
28. Bussemaker HJ, Foat BC, Ward LD (2007) Predictive modeling of genome-wide mrna expression: from modules to molecules. Annu Rev Biophys Biomol Struct 36: 329–47.
29. Park SJ, Nakai K (2011) A regression analysis of gene expression in es cells reveals two gene classes that are significantly different in epigenetic patterns. BMC Bioinformatics 12 Suppl 1: S50.
30. Irie T, Park SJ, Yamashita R, Seki M, Yada T, et al. (2011) Predicting promoter activities of primary human dna sequences. Nucleic Acids Res 39: e75.
31. Lauritzen S (1996) Graphical Models. New York: Oxford University Press.
32. Park SJ, Komata M, Inoue F, Yamada K, Nakai K, et al. (2013) Inferring the choreography of parental genomes during fertilization from ultralarge-scale whole-transcriptome analysis. Genes Dev 27: 2736–48.
33. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nat Protoc 7: 562–78.
34. Benz C, Copley MR, Kent DG, Wohrer S, Cortes A, et al. (2012) Hematopoietic stem cell subtypes expand differentially during development and display distinct lymphopoietic programs. Cell Stem Cell 10: 273–83.
35. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc 4: 44–57.
36. Umemoto T, Yamato M, Ishihara J, Shiratsuchi Y, Utsumi M, et al. (2012) Integrin-alphavbeta3 regulates thrombopoietin-mediated maintenance of hematopoietic stem cells. Blood 119: 83–94.
37. Domen J, Cheshier SH,Weissman IL (2000) The role of apoptosis in the regulation of hematopoietic stem cells: Overexpression of bcl-2 increases both their number and repopulation potential. J Exp Med 191: 253–64.
38. Peng C, Chen Y, Shan Y, Zhang H, Guo Z, et al. (2012) Lsk derived lsk- cells have a high apoptotic rate related to survival regulation of hematopoietic and leukemic stem cells. PLoS One 7: e38614.
39. Wingender E, Chen X, Hehl R, Karas H, Liebich I, et al. (2000) Transfac: an integrated system for gene expression regulation. Nucleic Acids Res 28: 316–9.
40. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106–17.
41. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008) Genome-scale dna methylation maps of pluripotent and differentiated cells. Nature 454: 766–70.
42. Moignard V, Macaulay IC, Swiers G, Buettner F, Schutte J, et al. (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat Cell Biol 15: 363–72.
43. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, et al. (2003) Match: A tool for searching transcription factor binding sites in dna sequences. Nucleic Acids Res 31: 3576–9.
44. Boyle P, Despres C (2010) Dual-function transcription factors and their entourage: unique and unifying themes governing two pathogenesis-related genes. Plant Signal Behav 5: 629–34.
45. Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, et al. (2012) Functional analysis of transcription factor binding sites in human promoters. Genome Biol 13: R50.
46. Diffner E, Beck D, Gudgin E, Thoms JA, Knezevic K, et al. (2013) Activity of a heptad of transcription factors is associated with stem cell programs and clinical outcome in acute myeloid leukemia. Blood 121: 2289–300.
47. Chute JP, Ross JR, McDonnell DP (2010) Minireview: Nuclear receptors, hematopoiesis, and stem cells. Mol Endocrinol 24: 1–10.
48. Henke BR, Blanchard SG, Brackeen MF, Brown KK, Cobb JE, et al. (1998) N-(2-benzoylphenyl)-l-tyrosine ppargamma agonists. 1. discovery of a novel series of potent antihyperglycemic and antihyperlipidemic agents. J Med Chem 41: 5020–36.
49. Brown KK, Henke BR, Blanchard SG, Cobb JE, Mook R, et al. (1999) A novel n-aryl tyrosine activator of peroxisome proliferator-activated receptor-gamma reverses the diabetic phenotype of the zucker diabetic fatty rat. Diabetes 48: 1415–24.
50. Liebermann DA, Gregory B, Hoffman B (1998) Ap-1 (fos/jun) transcription factors in hematopoietic differentiation and apoptosis. Int J Oncol 12: 685–700.
51. Shen LJ, Chen FY, Zhang Y, Cao LF, Kuang Y, et al. (2013) Mycn transgenic zebrafish model with the characterization of acute myeloid leukemia and altered hematopoiesis. PLoS One 8: e59070.
52. Yu S, Jing X, Colgan JD, Zhao DM, Xue HH (2012) Targeting tetramer-forming gabpbeta isoforms impairs self-renewal of hematopoietic and leukemic stem cells. Cell Stem Cell 11: 207–19.
53. Ghiaur G, Yegnasubramanian S, Perkins B, Gucwa JL, Gerber JM, et al. (2013) Regulation of human hematopoietic stem cell self-renewal by the microenvironment's control of retinoic acid signaling. Proc Natl Acad Sci U S A.
54. Okada S, Fukuda T, Inada K, Tokuhisa T (1999) Prolonged expression of c-fos suppresses cell cycle entry of dormant hematopoietic stem cells. Blood 93: 816–25.
55. Wozniak RJ, Keles S, Lugus JJ, Young KH, Boyer ME, et al. (2008) Molecular hallmarks of endogenous chromatin complexes containing master regulators of hematopoiesis. Mol Cell Biol 28: 6681–94.
56. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, et al. (2013) Charting a dynamic dna methylation landscape of the human genome. Nature 500: 477–81.
57. Kawana M, Lee ME, Quertermous EE, Quertermous T (1995) Cooperative interaction of gata-2 and ap1 regulates transcription of the endothelin-1 gene. Mol Cell Biol 15: 4225–31.
58. Zhang P, Behre G, Pan J, Iwama A, Wara-Aswapati N, et al. (1999) Negative cross-talk between hematopoietic regulators: Gata proteins repress pu.1. Proc Natl Acad Sci U S A 96: 8705–10.
59. Peixoto A, Monteiro M, Rocha B, Veiga-Fernandes H (2004) Quantification of multiple gene expression in individual cells. Genome Res 14: 1938–47.
60. Warren L, Bryder D, Weissman IL, Quake SR (2006) Transcription factor profiling in individual hematopoietic progenitors by digital rt-pcr. Proc Natl Acad Sci U S A 103: 17807–12.
61. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, et al. (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498: 236–40.
62. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol 10: R25.
63. Miranda-Saavedra D, De S, Trotter MW, Teichmann SA, Gottgens B (2009) Bloodexpress: a database of gene expression in mouse haematopoiesis. Nucleic Acids Res 37: D873–9.
64. Hu ZL, Bao J, Reecy J (2008) Categorizer: A web-based program to batch analyze gene ontology classification categories. Online J Bioinform 9: 108–12.

OPEN ACCESS Freely available online

PLOS | ONE

A Set of Structural Features Defines the *Cis*-Regulatory Modules of Antenna-Expressed Genes in *Drosophila melanogaster*

Yosvany López[1,2], Alexis Vandenbon[3], Kenta Nakai[2]*

1 Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan, **2** Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, **3** Immunology Frontier Research Center, Osaka University, Osaka, Japan

Abstract

Unraveling the biological information within the regulatory region (RR) of genes has become one of the major focuses of current genomic research. It has been hypothesized that RRs of co-expressed genes share similar architecture, but to the best of our knowledge, no studies have simultaneously examined multiple structural features, such as positioning of *cis*-regulatory elements relative to transcription start sites and to each other, and the order and orientation of regulatory motifs, to accurately describe overall *cis*-regulatory structure. In our work we present an improved computational method that builds a feature collection based on all of these structural features. We demonstrate the utility of this approach by modeling the *cis*-regulatory modules of antenna-expressed genes in *Drosophila melanogaster*. Six potential antenna-related motifs were predicted initially, including three that appeared to be novel. A feature set was created with the predicted motifs, where a correlation-based filter was used to remove irrelevant features, and a genetic algorithm was designed to optimize the feature set. Finally, a set of eight highly informative structural features was obtained for the RRs of antenna-expressed genes, achieving an area under the curve of 0.841. We used these features to score all *D. melanogaster* RRs for potentially unknown antenna-expressed genes sharing a similar regulatory structure. Validation of our predictions with an independent RNA sequencing dataset showed that 76.7% of genes with high scoring RRs were expressed in antenna. In addition, we found that the structural features we identified are highly conserved in RRs of orthologs in other *Drosophila* sibling species. This approach to identify tissue-specific regulatory structures showed comparable performance to previous approaches, but also uncovered additional interesting features because it also considered the order and orientation of motifs.

Citation: López Y, Vandenbon A, Nakai K (2014) A Set of Structural Features Defines the *Cis*-Regulatory Modules of Antenna-Expressed Genes in *Drosophila melanogaster*. PLoS ONE 9(8): e104342. doi:10.1371/journal.pone.0104342

Editor: Miguel A. Andrade-Navarro, Max Delbrück Center for Molecular Medicine, Germany

Received February 7, 2014; **Accepted** July 13, 2014; **Published** August 25, 2014

Funding: This work was partly supported by Grants-in-Aid for Scientific Research from JSPS (25290067). Yosvany Lopez thanks the support of the MEXT scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: knakai@ims.u-tokyo.ac.jp

Introduction

Understanding the biological information encoded in RRs of genes constitutes one of the greatest challenges in genomics. Analysis of regulatory structure can provide important insight into interactions with specific transcription factors (TFs) and can help predict genes that will be expressed in certain tissues, cell types, or physiological conditions.

Several recent studies have revealed interesting details about regulatory structure and TF binding sites (TFBSs). *Cis*-regulatory elements and motif pairs that are bound by interacting proteins have demonstrated the co-occurrence of specific TFBS in some promoters [1]. Many of the genomic regions that are densely bound by TFs have also revealed new binding relationships between factors [2]. Other studies have examined dependencies among TFBSs. For instance, a set of rules to define the presence and pairwise positioning effects of motifs was developed for modeling human and mouse promoters [3]. Novel motif patterns have also been observed in the promoters of co-expressed genes in *Arabidopsis thaliana* [4]. However, none of these studies considered the orientation, pairwise positioning, and order of the motifs within the RR.

Antenna is a sensory organ located in the anterior part of an insect's head. It is usually covered with olfactory receptors able to detect odor particles in the air, and is sometimes used as humidity sensors for detecting changes in vapor water concentrations. The function of antenna has been studied for understanding the receptor-odorant interactions [5] and analyzing the expression profiles of odorant binding proteins [6]. Other studies have also addressed how flies use the sensing of air motion for controlling flight [7]. Given the importance of antenna and that *Drosophila melanogaster* is a well-studied model organism with a large amount of available genomic data to validate new findings, we have chosen the co-expressed genes in *D. melanogaster* antenna for our analysis of RRs.

Quantitative analyses of enhancer activity of different DNA sequences have revealed many cell type-specific *D. melanogaster* enhancer sequence elements [8]. Computational approaches for finding *cis*-regulatory modules using thermodynamic modeling based on *D. melanogaster* TFBS preferences suggest that positional information is highly important and that weak and strong TFBSs contribute equally to regulation of gene expression [9]. Furthermore, a machine-learning framework that integrated TF binding,

evolutionarily conserved sequence motifs, gene expression and chromatin modification data was designed to predict putative functions for uncharacterized genes involved in *D. melanogaster* nervous system development [10]. This integrated framework demonstrated a complementarity between physical evidence of regulatory interactions and coordinated expression. Similarly, reporter gene assays have demonstrated organ-specific expression patterns in *D. melanogaster* [11].

Although some solitary TFBSs are potentially functional, most methods intended to identify *cis*-regulatory modules do not take them into consideration. In addition, despite the clear interdependency among TFBSs, no computational method has simultaneously examined positional and structural relationships of different motifs to model the RR of co-expressed genes. In general, details about the regulatory structures responsible for regulating tissue- or condition-specific gene expression are still lacking.

Here we report a novel computational method that incorporates several different structural features (SFs), including motif orientation, order, position relative to the transcription start site (TSS), and pairwise positioning of motifs to wholly describe the regulatory architecture of *D. melanogaster* antenna-expressed genes. Although a previous framework combined some of these SFs [12], it did not consider the order of regulatory motifs, focusing instead on motif discovery.

Since a broad genomic region around the TSS is considered in this work, we will analyze the *cis*-regulatory modules of *Drosophila* genes, rather than only their core promoter region. To avoid confusion, we will define the "regulatory regions" (RR), as regions that comprise not only the *Drosophila* core promoter region but also enhancers located in its proximity.

This analysis initially predicted six motifs in the RR of antenna-expressed genes, three of which appeared to be novel. We then created a feature collection using all of the predicted motifs and removed irrelevant features with a correlation-based filter. The resulting feature set was further optimized with a genetic algorithm (GA), which achieved an area under the curve (AUC) of 0.841 and produced eight features that best characterize the RR of antenna-expressed genes. This final feature set was used to score all the RRs of *D. melanogaster* genes for unknown antenna-expressed genes sharing a similar regulatory structure. Of the 1000 genes with the highest-scoring RRs, RNA sequencing (RNA-seq) data of antenna-related cell types showed that 76.7% of them were expressed in antenna tissue. We next searched for the presence of our SFs across the *Drosophila* lineage and found evidence for their conservation in the RR of orthologs in other sibling species. Finally, we also used a set of *Caenorhabditis elegans* muscle-expressed genes to compare our method to a similar approach [13], thus uncovering relevant SFs related to the order and orientation of regulatory motifs.

Results

Our approach consisted of three main steps (Figure 1): the first step focused on identifying over-represented motifs in the RR of antenna-expressed genes, the second step focused on generating a broad set of SFs and the third step is intended to optimize the set of SFs that best describe the RR of these genes. Co-expression data were obtained for an initial set of 224 *D. melanogaster* antenna-expressed genes from COXPRESdb [14]. The initial set was randomly split into three exclusive subsets: a "motif-prediction" set (90 genes), a "feature-generation" set (44 genes), and a "model-build" set (90 genes).

Predicted antenna-related motifs

We first predicted *cis*-regulatory motifs in the 90 RRs (1.5 kbp upstream and 500 bp downstream of the TSS) of antenna-expressed genes in the "motif-prediction" set. We initially uncovered 65 *de novo* motifs. After removal of redundancy in this motif set, 25 non-redundant motifs remained. By using the same "motif-prediction" set, we computed the over-representation index (ORI) [15] for these motifs and removed those with low levels of enrichment in the RRs of antenna-expressed genes. Thus, our final motif collection contained six highly enriched, non-redundant motifs, which we designated *D. melanogaster* enriched (DME) 1–6 (Figure 2). These motifs were compared with those in the JASPAR CORE Insecta database of eukaryotic TF binding profiles [16] and three significant matches were found. DME-4 matched the motif bound by the TFs Eip74EF (ecdysone-induced protein 74EF) and STAT92E (signal transducer and transcription activator), whereas DME-5 and DME-6 matched the motifs bound by the TFs Eip74EF and opa (pair-rule protein odd-paired). To the best of our knowledge, none of these motifs has been reported to be important in antenna. The analysis of acetylation patterns on *Drosophila* ecdysone induced Eip74EF and Eip75B genes has shown acetylation of histone H3 lysine 23 in promoters and its relationship to ecdysone induced gene activation [17]. The activation of STAT92E, a signal transducer in early wing imaginal discs has been shown to inhibit the formation of ectopic wing fields whereas specifies dorsal pleural and inhibits notum identity to divide the body wall [18]. The TF opa1, on the other hand, increases mitochondrial morphometric heterogeneity, thus allowing heart dilation and contractile impairment in *Drosophila* [19]. For the remaining three motifs we did not find any significant match in the JASPAR CORE Insecta database [16], so they appear to be new motifs with potentially important roles in regulating antenna-expressed genes. Comparisons of our six motifs with other previously found in *Drosophila* [20] showed a certain similarity of motifs DME-3 and DME-6 to Motif 7 and Motif 1 (see Table 2 in [20]), respectively.

Generated and filtered SFs

The six over-represented motifs were used to scan the RRs of genes in the "feature-generation" set for SFs based on position relative to the TSS, pairwise positioning, orientation, and order of these motifs. The regions were scanned in 100-bp windows in both directions (1.5 kbp upstream and 500 bp downstream) from the TSS (Figure 3), and we identified 544 features. To describe the order of motifs, the positions of no more than three motifs were considered per feature. We binarized the features so that each RR was represented as a vector where the presence (1) or absence (0) of each feature was indicated. The 544 features were also examined in the RRs of genes in a negative control set (genes with low expression in antenna; Z-score <-1). We then built a 544×1117 binary matrix (544 features; 44 genes in the "feature-generation" set and 1073 genes in the negative control set). This feature set was filtered with a correlation-based filter [21], which removed any features for which the correlation with the RR of genes in the "feature-generation" set did not predominate, even after removing redundant features. After filtering, 19 SFs remained.

Optimization of the SFs

We weighted the 19 filtered features based on their relevance in describing the RR of antenna-expressed genes and designed a GA to obtain the most informative combination of features. Unlike traditional machine learning methods, GA operates without *a priori* knowledge of the problem to be solved. When used in optimization problems, they tend to be less affected by local

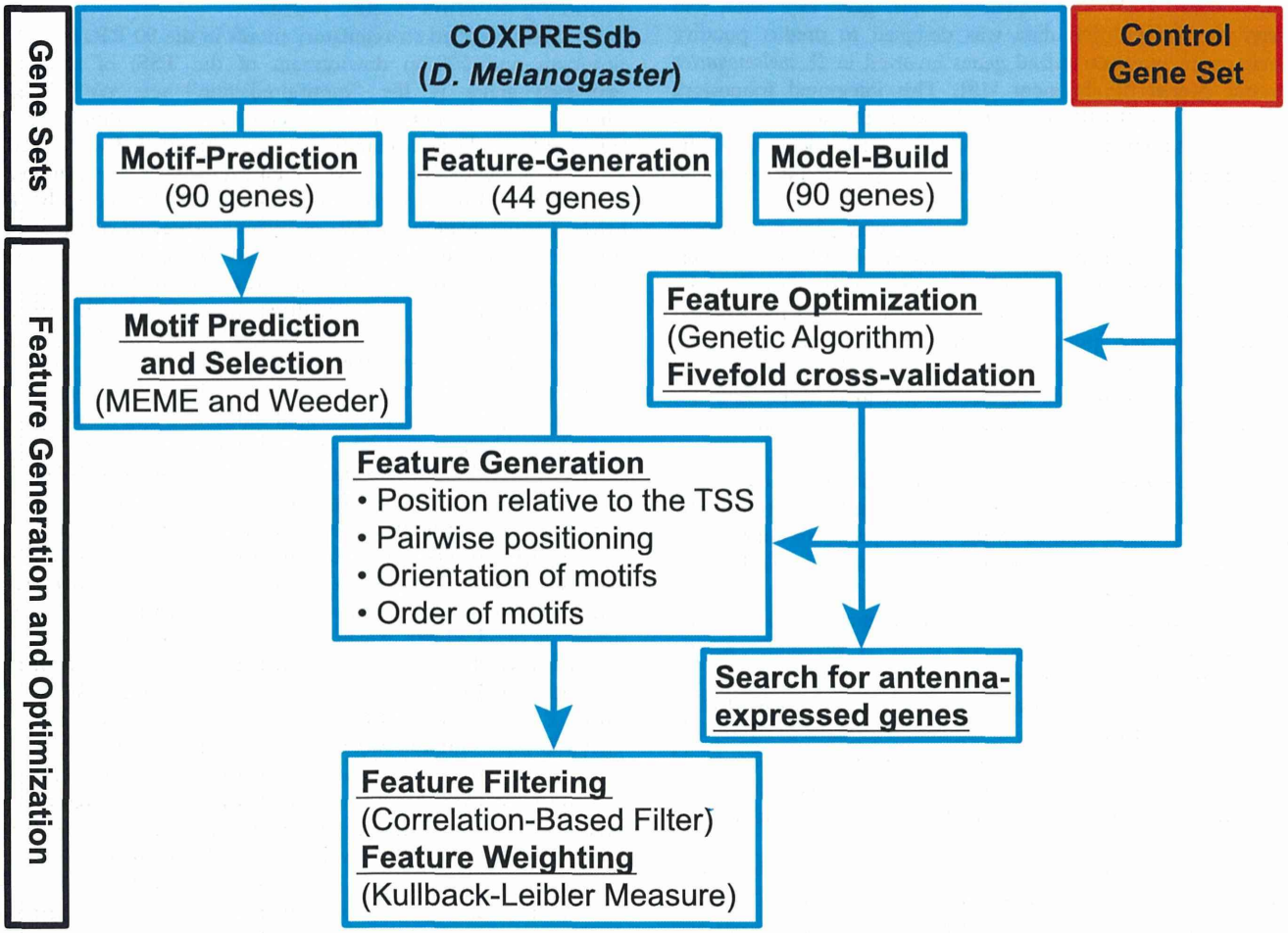

Figure 1. Workflow of our computational method.
doi:10.1371/journal.pone.0104342.g001

maxima than other methods. Because of these advantages, we employed a GA to identify the best combination of features. The "model-build" and negative control gene sets were randomly split into five subgroups, and the GA was trained with four of the subgroups and tested with the remaining one. The fivefold cross-validation (CV) method [22] was repeated 100 times, and the best CV run of the GA, which achieved an AUC of 0.841 (Figure 4), was considered for further analysis. After this validation process, the previous collection of 19 features was reduced to eight high-confidence SFs (Figure 5).

Genes sharing similar regulatory structure

To evaluate the biological validity of the eight identified SFs, we used them to scan the entire *D. melanogaster* genome for genes with a similar regulatory structure. By using a scoring system that sums up the weight of every present SF, we scored each RR according to the SFs it contained and selected the 1000 genes with the highest-scoring regions. We next obtained the Gene Ontology (GO) terms [23] (uncorrected p-value ≤ 0.01) for these genes and found that a reduced subset of them appear to function in "bristle morphogenesis", in the biological process that generates sensory bristle structures, or in basal functions of the cell (Table S1). Because the corrected GO term p-values were exceptionally high, probably owing to the lack of complete annotation data, we further mapped the RNA-seq data of two cell lines in the third instar larval stage to *D. melanogaster* genome. The cell lines were taken from the tissue eye-antenna disc-derived cell-line (DCCid: modENCODE_4399) and antenna disc-derived cell-line (DCCid: modENCODE_4402), respectively. We found that 7,691 (63.1%) of 12,192 genes in the genome-wide set were expressed in antenna, whereas 767 (76.7%) of 1000 genes with high-scoring RRs according to our identified features were expressed in the antenna-related cell types. From the 7,691 antenna-expressed genes, 5,666 of them were among the 7,691 genes with highest-scoring RRs. This percentage of antenna-expressed genes (76.7%) is given because we have used a high threshold (Fragments Per Kilobase of transcript per Million mapped reads (FPKM)>1) compared to previous studies [24] (FPKM>0.05). Because this expression data originated from immature cells, many receptor genes showed little or no expression at all.

We noted that among the 50 genes with highest-scoring RRs (Figure S1 and Table S2), only two were also included in the "motif-prediction", "feature-generation", and "model-build" sets. Since each gene in the initial sets has different SFs, genes with RRs containing more SFs or more heavily weighted SFs will score higher compared to others. From the initial set of 224 genes, 81 genes were among the 1000 top scoring genes. We next verified how many of the 50 highest-scoring RRs contained the identified SFs. We found that all 50 of the RRs contained DME-3 at ~0–100 bp from DME-3 on the plus strand (feature 1), 11 RRs had DME-5 at ~100–200 bp from the TSS on the minus strand (feature 2), 34 RRs had DME-4 at ~200–300 bp from the TSS on

ID	Logo	ORI	Comment	Citations
DME-1		2.27	-	-
DME-2		3.09	-	-
DME-3		2.19	-	-
DME-4		2.24	Eip74EF (5.82e-04) STAT92E (2.91e-03)	[17, 18]
DME-5		2.08	Eip74EF (1.14e-04)	[17]
DME-6		2.40	opa (3.32e-03)	[19]

Figure 2. Predicted motifs in RRs of antenna-expressed genes. For each motif, the identifier, logo, and ORI are shown. The known regulatory motif, TOMTOM *p*-value, and citations are also given for motifs that matched already identified motifs.
doi:10.1371/journal.pone.0104342.g002

either strand (feature 3), 40 RRs had DME-5 at ~600–700 bp from the TSS on either strand (feature 5), and 19 RRs had DME-6 at ~300–400 bp from the TSS on either strand (feature 8). The scoring of *D. melanogaster* RRs uncovered genes with known biological functions in sensory organs and others with unknown biological function. Figure 6 depicts four of the 50 highest-scoring RRs, three of which are involved in detecting chemical stimuli, sensory organ development, and neurogenesis, and one with unknown biological function. *Gr22b* (FlyBase ID FBGN0045500) encodes a protein involved in detecting chemical stimuli [25]. The RR of *Gr22b* shares three SFs 1, 3, and 5 with that of *ac* (FlyBase ID FBGN0000022) and *Adk2* (FlyBase ID FBGN0022708), which encode proteins involved in sensory organ development and neurogenesis [26,27]. The RR of gene CG17298 (FlyBase ID FBGN0038879) shares the previous three features with that of genes *Gr22b*, *ac* and *Adk2* while also containing feature 8 (Figure 6).

Furthermore, genomes of 11 *Drosophila* sibling species were downloaded from FlyBase database [28]. RRs of each *Drosophila* specie's genes were extracted. Each RR was scanned for potential binding sites of the six enriched antenna-related motifs. We next scanned every RR for the presence of our eight SFs. As a result, we found that feature 1 is extensively conserved across *Drosophila* orthologs. In addition, the RRs of the closest orthologs mostly share features 2, 3 and 5 (Figure 7 and Figures S2 and S3).

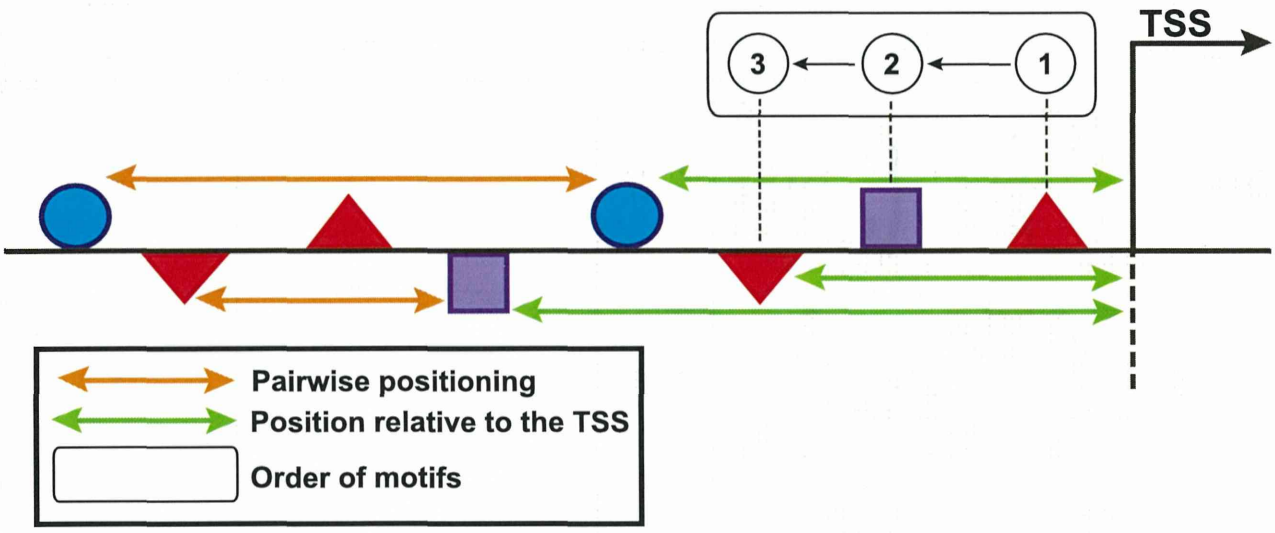

Figure 3. Schematic of our upstream RR-scanning approach. The same approach was followed for the downstream regions. The geometrical forms on and under the black line represent the TFBS on the plus and minus strand, respectively. The orange and green lines and the rectangle indicate the computed SFs.
doi:10.1371/journal.pone.0104342.g003

Comparison with another method

We compared our computational method to a similar reported promoter structure-modeling approach [13] with a set of 121 *C. elegans* muscle-expressed genes. We randomly split such a set into three independent subsets: a "Ce motif-prediction" set (48 genes), a "Ce feature-generation" set (23 genes), and a "Ce model-build" set (50 genes). The *C. elegans* genome was obtained from WormBase [29]. RR extending from 1 kbp upstream to 200 bp downstream of the TSS was analyzed. Two different motif-discovering algorithms: MEME [30] and Weeder [31] were used for predicting *de novo* motifs in the RRs of muscle-expressed genes in the "Ce motif-prediction" set. A total of 64 *de novo* motifs were uncovered, and 18 non-redundant motifs were obtained after removing redundancy. We next computed the ORI [15] of each previous motif, leaving us with 11 over-represented motifs (Table S3). Comparison of the motifs with those in the JASPAR CORE

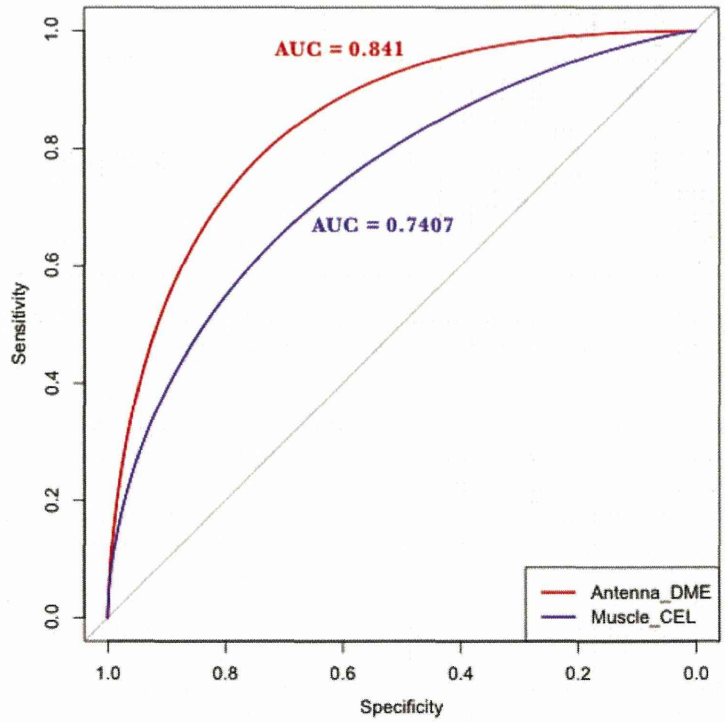

Figure 4. Performance of our GA with two different sets of co-expressed genes. The red line represents the AUC for antenna-expressed genes in *D. melanogaster*, and the blue line represents the AUC for muscle-expressed genes in *C. elegans*.
doi:10.1371/journal.pone.0104342.g004

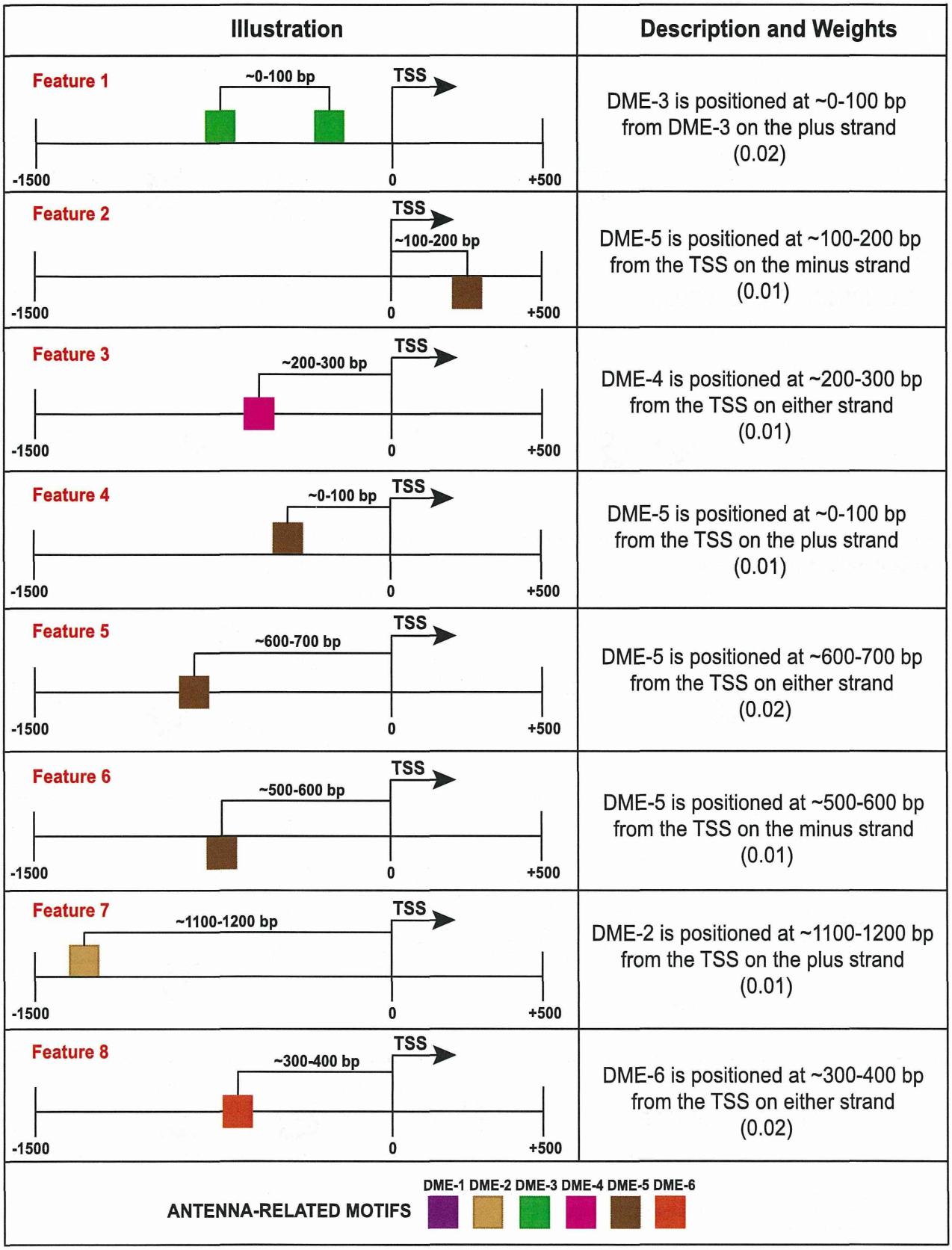

Figure 5. The set of SFs that best describe the RRs of antenna-expressed genes in *D. melanogaster*. For each feature, the relationship between motifs within the feature and the Kullback-Leibler weight are shown. The colored squares represent the antenna-related motifs. Squares on or under the black line indicate motifs on the plus or minus strand, whereas squares in the middle of the black line indicate motifs on either strand.
doi:10.1371/journal.pone.0104342.g005