Estimating the significance of Average Motif Affinity scores

The Average Motif Affinity (AMA) score of a DNA sequence represents the total binding affinity a TF. It is defined as the average likelihood ratio of all positions (sites) on the sequence. The likelihood ratio of a site is the probability of the site under a motif model, M, divided by the probability of the site under a background model, B. We represent M as a position specific probability matrix, and B by the parameters of a zero-order sequence model of DNA. The definition of the AMA score is given in Eqn. 1.

$$\text{AMA}(X, M) = \frac{1}{N} \sum_{site \in X} \frac{Pr(site|M)}{Pr(site|B)}, \quad (1)$$

We assign AMA scores to sequences, and would like to estimate the p-value of each score. This requires that we somehow estimate the null distribution of the scores. We would like the p-value assigned to a sequence to be based on a permuted-letter null model. We could estimate this null score distribution directly by multiply permuting the sequence and computing the motif-based score. This is computationally expensive and results in p-values censored below $1/N$, where N is the number of permutations we carry out.

To avoid the two limitations of permutation-based p-value estimation, we take a dynamic programming approach. Our approach first computes the score distribution of the likelihood ratio score of a single site. This score distribution assumes two distinct zero-order random sequence models: B, the model used for computing the likelihood ratio of a site, and B', the model for a random (e.g., permuted) sequence. Our approach then uses a second round of dynamic programming to estimate the distribution of the average of n scores under the simplifying assumption that the scores of overlapping sites are independent. In order to reduce computation time, we only compute the score distribution for certain values of n and for certain values of B'. To assign a p-value to a sequence, we use interpolation on n and B' based on its length and base composition. The details of our p-value estimation approach are described in what follows.

To estimate the score distribution of the likelihood ratio of a site, we adapt a standard approach often used for log likelihood ratio scores (Bailey and Gribskov, 1998). We first convert M to a position-specific log likelihood matrix, M', by dividing each entry in M by the corresponding value in B and taking logs,

$$S_{a,i} = \log_2 \frac{M_{a,i}}{B_a}, a \in \{A, C, G, T\}, i \in \{1, \ldots, w\},$$

where w is the motif width. We scale and round the entries to be in the range $[0, \ldots, 100]$ and call this scoring matrix S'. Dynamic programming is then used to successively compute the score distribution for the first i columns of the motif, T_i for $i = 1, \ldots, w$. If we define $T_{i,x} = Pr(T_i = x)$, then the recursion

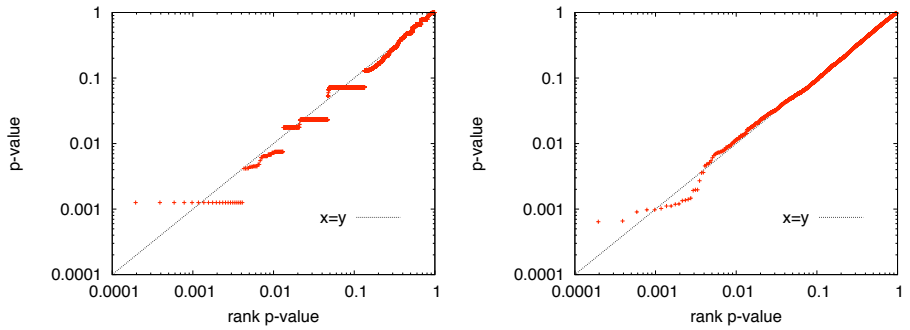

Figure 1: AMA *p*-**values of permuted yeast promoter sequences.** The figure shows Q-Q plots of *p*-values for a synthetic motif used to scan permuted yeast promoter sequences. Panel **a)** shows results for *p*-values without GC-correction. Panel **b)** shows results for *p*-values with GC-correction.

formulas are

$$T_{i,x} = \begin{cases} \sum_{\{a|x=S'_{a,i}\}} B'_a & \text{if } i = 1, \\ \sum_{t=0}^{100i} \sum_{\{a|x=t+S'_{a,i}\}} T_{i-1,t} B'_a & \text{for } i \in \{2,\ldots,w\}. \end{cases}$$

The vector $T_{w,x}$ contains the distribution for score x, where x is the (scaled) logarithm of the AMA score for a sequence of length w, the width of the motif. Since exactly one site will fit in a sequence of length w, we define $T_x^{(1)} = T_{w,x}$. To compute the score distribution for longer sequences, we use convolution. By convolving $T_x^{(1)}$ with itself, we get the distribution for sequences where two sites will fit, $T_x^{(2)}$. Note that during convolution we convert the scaled, log scores, x, to AMA scores, take their average, and then reverse the conversion. Convolving the $T_x^{(2)}$ distribution with itself gives us $T_x^{(4)}$. Repeating this trick allows us to create vectors containing the distributions for $n = 1, 2, 4, \ldots, L = 2^j$ in time proportional to $log_2(L)$. When computing the *p*-value of a sequence whose length is not a power of two, $2^{j-1} < L < 2^j$, we interpolate between the *p*-values corresponding to sequences of length 2^{j-1} and 2^j, as we describe below.

We have now constructed a lookup table for *p*-values assuming a particular sequence model, B'. In order to estimate *p*-values for sequences with different base distributions, we repeat the above steps with different values of B'. In particular, we construct *p*-value lookup tables for "balanced" distributions $B' = f_A, f_C, f_G, f_T$, where $f_C = f_G$ and $f_A = f_T$. Such distributions can be specified by a single parameter, e.g. the frequency of G, f_G. We choose a number (typically about 40) of equally spaced values of f_G, and estimate their corresponding score distributions as described above. Using the resulting lookup tables, we use bi-linear interpolate on f_G and $log(2L)$ to obtain the *p*-value for the score of a sequence of length L with a GC content of $2 * f_G$. The extra

factor of two in the length value is due the fact that AMA scores are actually the average over sites on both DNA strands.

Importance of GC-compensation

The importance of using GC-compensated p-values is illustrated in Fig. 1, which shows typical results using yeast TF motifs. The figure shows Q-Q plots for a single motif used to scan all yeast promoters, where the promoters have been permuted to simulate null data. Without GC-compensation (Fig. 1**a**), the estimated p-values (Y-axis) tend to be smaller than expected (rank p-values, X-axis). With GC-compensation, the p-values correlated extremely well with the expected (rank) p-values.

Comparison with HMM0 p-values

We compare the p-values calculation by AMA (analytical) and HMM0 (empirical) both with respect to runtime and correlation. For HMM0, empirical p-values are calculated based on shuffled promoter sequences (100x) with parameter values set according to the literature (Sinha *et al.*, 2008). The experiment was performed on a 2.66 GHz Intel Xeon processor. Both methods are used to score the yeast promoters (750 bp) using all 124 yeast TF motifs from our study. Fig. 2**a** shows the average runtime in seconds used by either of the programs for a particular motif size. For motifs of width less than 10, AMA with and without GC-compensation is at least a magnitude faster than HMM0. The difference in runtime becomes less significant with motif widths above 10 when calculating GC-compensated p-values in AMA.

We previously reported that sequence composition can bias GOMO's prediction (Bodén and Bailey, 2008). In this study we therefore use GC-compensated p-values. The histogram in Fig. 2**b** shows the correlation coefficient of the p-values calculated by AMA and HMM0 for each of the 124 motifs in *S. cerevisiae*. For most of the motifs the correlation is very high ($cc \geq 0.85$).

We find that the correlation declines with increasing motif width. Some of the longest motifs—ARO80 (23), DAL81 (19), YOX1 (20), SNF1 (17) and STP4 (15)—show correlations of 0.1, 0.05, 0.26, 0.09, and 0.45 respectively. The lower correlation observed for p-values of wide motifs may be due to differences in the underlying scoring schemes. Whereas AMA averages the affinity of all possible sites in a sequence, HMM0 averages the affinity only over all possible configurations of non-overlapping sites.

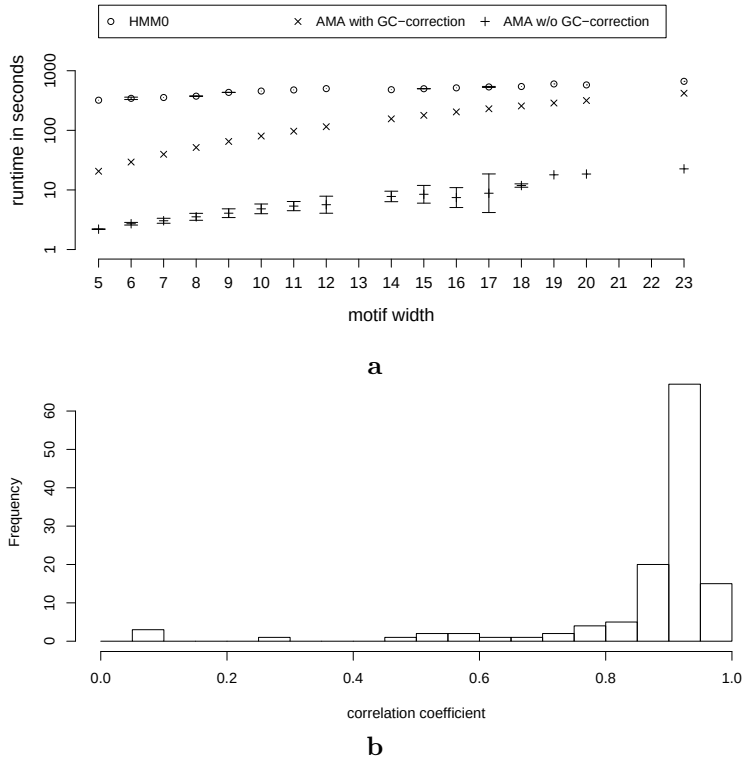

Figure 2: **Comparison between** AMA **and** HMM0**. a)** Run-time of AMA and HMM0. Each point shows the average run-time in seconds (Y) of the AMA or HMM0 algorithms using motifs of the given width (X). **b)** Correlation between p-values calculated by AMA and HMM0.

Comparison with alignment-based method

We replaced AMA scores with scores derived from MONKEY (Moses *et al.*, 2004) p-values and evaluated GOMO on the yeast multiple-species datasets. We tested two functions for scoring a single promoter: 1) the minimum p-value; 2) the geometric mean of all p-values. MONKEY was used to compute the p-value of the match of each position in the promoter to the given TF binding motif. For each promoter, MONKEY searched a multiple alignment of the upstream regions of all four yeast orthologs. MONKEY used the species tree derived from all yeast intergenic regions by Kellis *et al.* (2003), the "HB" model and *S. cerevisiae* was used as the "key" species by MONKEY in its heuristics. All other MONKEY parameters were defaults. The multiple alignment for each promoter was made

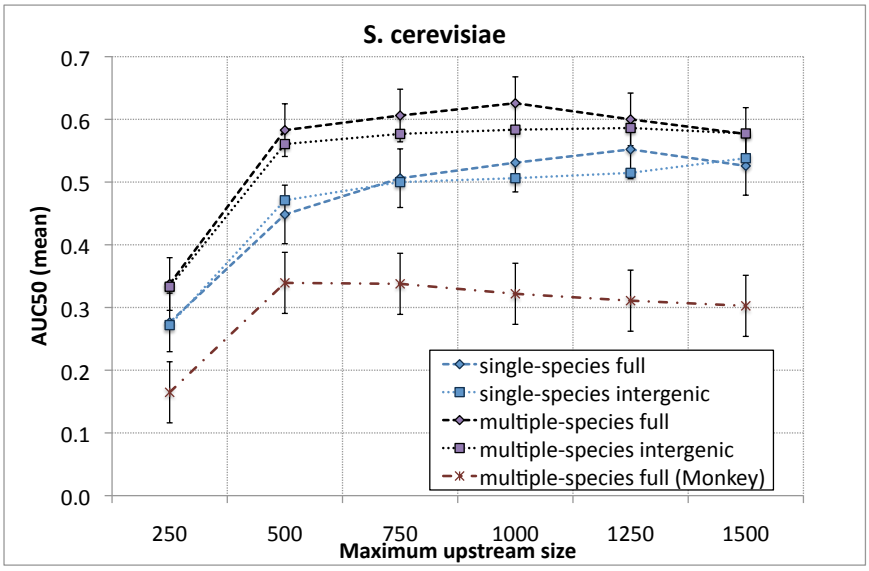

Figure 3: **Multiple-species** GOMO **prediction accuracy.** Each point shows the average AUC50 of TF-GO term association predictions made by GOMO in the key species *S. cerevisiae*. Points labeled "multiple-species" are results using promoter sequences from the key species and three related species; Monkey (Moses *et al.*, 2004) results use Monkey minimum p-value scores instead of AMA scores. Points labeled "single-species" are results using promoter sequences from the key species only, and are shown for comparison. The AUC50 is computed using a single TF, then averaged over TFs. The X-axis shows the upstream extent of promoter sequences ("full"), or the maximum upstream extent when they are truncated at the first ORF ("intergenic"). For clarity, standard error bars are shown for the "full" promoter sequence set only; standard error bars for the "intergenic" promoter set are similar.

using CLUSTALW2 (Chenna *et al.*, 2003) with all default settings applied to the same set of orthologous sequences as used by AMA. MONKEY-derived scores (derived from multiple alignments) were then input to single-species GOMO.

As seen in Fig. 3, GOMO achieves substantially lower accuracy using the minimum MONKEY p-value score instead of the AMA score for ranking promoters. The AUC50 for all sizes of upstream regions considered is less than half that achieved using the AMA score with multiple-species GOMO. The performance of the geometric mean MONKEY p-value score is even worse (data not shown).

Somewhat surprisingly, the minimum MONKEY p-value scoring function performs more poorly with GOMO than does the single-species AMA score. This may be due to several factors. Firstly, MONKEY is designed to predict individual binding sites, not target genes, and better ways to combine MONKEY scores for predicting target genes may exist, although we are not aware of them.

Secondly, like all alignment-based algorithms, MONKEY predictions are only as good as the alignments on which they are based, and multiple alignment is a notoriously hard problem. Thirdly, MONKEY predictions assume that the positions of binding sites are conserved, but much research indicates that binding sites frequently "drift" (Moses *et al.*, 2006). Combining AMA scores, which are specifically designed to predict TF target genes, and our method of combining GO-TF association scores in multiple-species GOMO, circumvents each of these three issues.

Proportion of correct predictions at relaxed false discovery rates

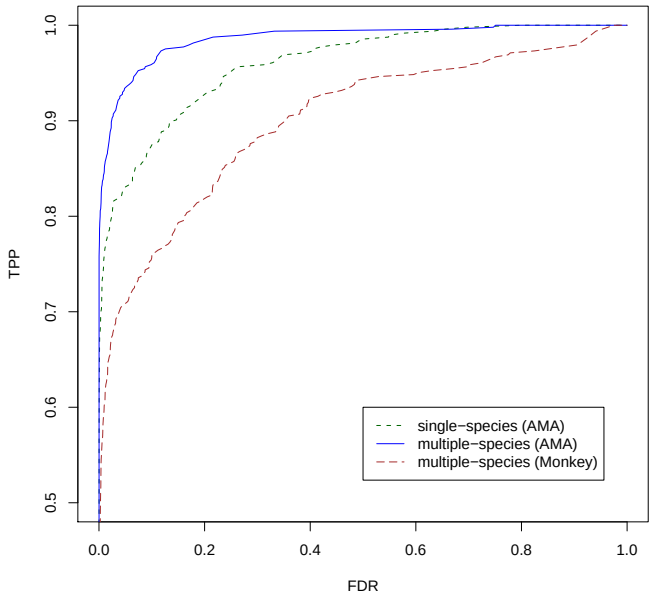

Figure 4: **Multiple-species GOMO prediction accuracy in yeast: sensitivity vs. FDR.** ROC-like curves for three different versions of GOMO applied to yeast 1000bp "full" promoter regions are shown. Predictions using all 42 yeast motifs contained in the gold standard are pooled, and true positive proportion is plotted as a function of the false discovery rate. The curve labeled "multiple-species (Monkey)" corresponds to using minimum p-value Monkey-based scores as input to single-species GOMO.

In order to compare the coverage of single- and multiple-species GOMO predictions at different FDR values, we create an ROC-like plot. As in an ROC plot, we plot the true positive proportion ($TP/(TP+FN)$) on the Y-axis. On

the X-axis, rather than false positive proportion ($FP/(TN+FP)$), we plot the false discovery rate ($FP/(TP+FP)$). In order to plot a single ROC-like curve for each type of prediction algorithm, we pool all predictions made for all motifs contained in our yeast "gold standard", and we use the 1000bp "full" promoter sequence datasets. We note that such plots allow comparison of the expected sensitivities of the prediction algorithms at any given false discovery rate.

Figure 4 shows the ROC-like plots for single- and multiple-species GOMO using the AMA scoring function, and for GOMO using the Monkey (Moses *et al.*, 2006) multiple-alignment-based scoring function (minimum p-value scores). The sensitivity of multiple-species GOMO dominates that of the other two algorithms at all false discovery rates. The relatively poor performance of the Monkey-based scores is discussed in the previous section (Comparison with alignment-based method).

GOMO pipeline

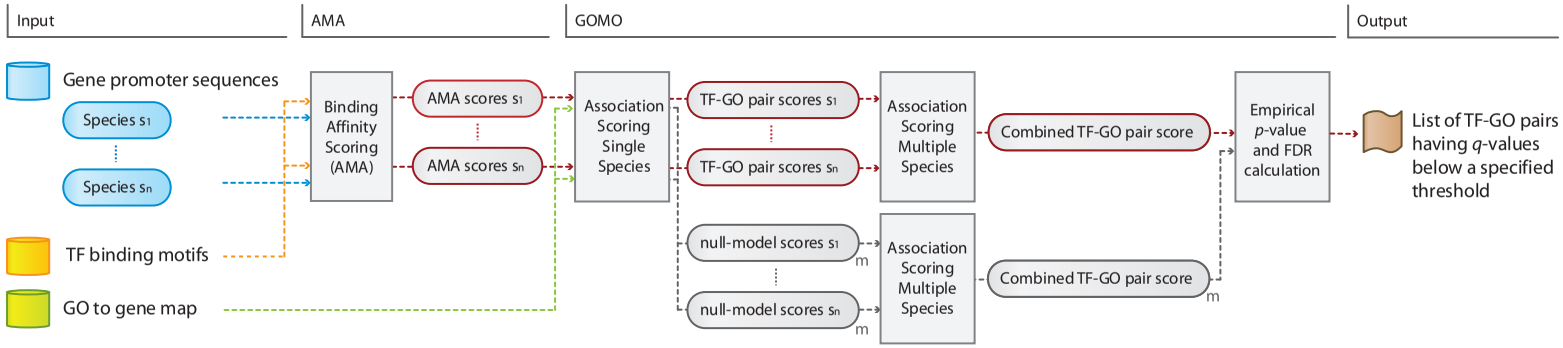

Figure 5: **Cartoon of the** GOMO **pipeline.** In order to calculate the minimum false discovery rate (q-value) one hundred independent sets of null scores are generated by randomly shuffling the gene ids, repeating the steps shown in gray for $m = [1, \ldots, 100]$.

[h]

Role-centric regulatory map for *H. sapiens* NFKB1

To explore the coherence of the roles predicted by multiple-species Gomo in *H. sapiens*, and to illustrate one use of such predictions, we create a role-centric regulatory map using Cytoscape Killcoyne *et al.* (2009) (Supplementary File 3 and Supplementary File 4). The nodes in the map are *H. sapiens* TFs and GO terms, and links indicate the most-specific predicted GO term of a TF. As an example, we explore the relationship of NFKB1 to other TFs by extracting the portion of the map containing NFKB1 and its nearest "neighbors"—all TFs that share a most-specific predicted GO term with NFKB1 (Fig. 6).

The NFKB1 role-centric regulatory map contains over 50 TF role predictions made by multiple-species Gomo, and identifies 14 "neighbor" TFs. Almost all of the role predictions for NFKB1 made by Gomo involve immune response, an important known function of this TF. Five of NFKB1's neighbor TFs are connected to it via immune response-related GO terms, and are also important regulators of immune response: REL, RELA, TLX1-NFIC, and SPIB, TAL1-TCF3 Hayden and Ghosh (2004); Hoffmann *et al.* (2004); Schotte *et al.* (2003); Voronova and Lee (1994); Kim *et al.* (2002); Palomero *et al.* (2006). One of these neighbor TFs—TAL1-TCF3—is known to directly regulate NFKB1 Chang *et al.* (2006). Two others, REL and RELA, which are known to form heterodimers with NFKB1 Hayden and Ghosh (2004), are also linked to NFKB1 in the map via the GO term "I-kappaB/NF-kappa complex". The remaining nine neighbor TFs are connected to NFKB1 via very general GO terms such as "system process" and "extracellular space". Interestingly, three of these nine TFs are known to share regulatory targets with NFKB1 and to be active during immune response: GATA-2 Ferla *et al.* (2002), SRF Pierce *et al.* (1995) and YY1 Tone *et al.* (2007). Although anecdotal, these results illustrate that the TF role predictions made by multiple-species Gomo can reveal biologically important relationships among *H. sapiens* TFs, in addition to discovering their individual biological roles.

References

Bailey, T. L. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**(1), 48–54.

Bodén, M. and Bailey, T. L. (2008). Associating transcription factor-binding site motifs with target GO terms and target genes. *Nucleic Acids Res*, **36**(12), 4108–4117.

Chang, P.-Y., Draheim, K., Kelliher, M. A., and Miyamoto, S. (2006). Nfkb1 is a direct target of the tal1 oncoprotein in human t leukemia cells. *Cancer Res*, **66**(12), 6008–6013.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, **31**(13), 3497–3500.

Ferla, K. L., Reimann, C., Jelkmann, W., and Hellwig-Bürgel, T. (2002). Inhibition of erythropoietin gene expression signaling involves the transcription factors gata-2 and nf-kappab. *FASEB J*, **16**(13), 1811–1813.

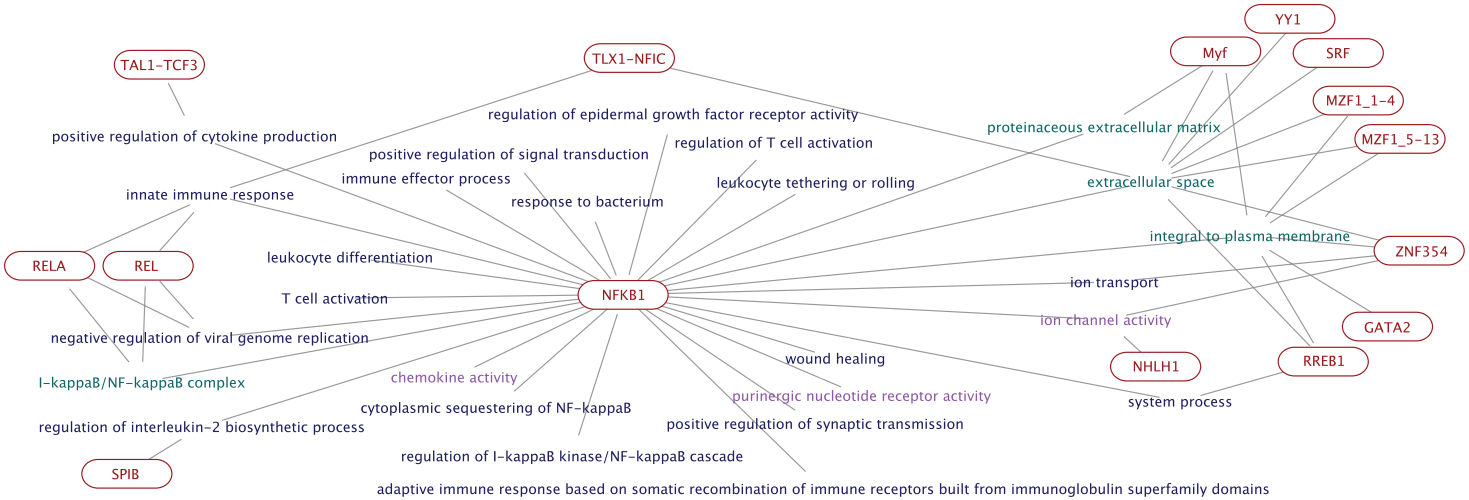

Figure 6: **Role-centric regulatory map for NFKB1.** The figure shows all the transcription factors (red boxes) that are predicted by multi-species GOMO to be "neighbors" in terms of their biological roles, and all of NFKB1's most-specific predicted GO terms (biological process - blue, molecular function - purple, cellular compartment - cyan). Two TFs are defined to be "neighbors" if they share a most-specific GO term predicted by GOMO.

Hayden, M. S. and Ghosh, S. (2004). Signaling to nf-kappab. *Genes Dev*, **18**(18), 2195–2224.

Hoffmann, K., Dixon, D. N., Greene, W. K., Ford, J., Taplin, R., and Kees, U. R. (2004). A microarray model system identifies potential new target genes of the proto-oncogene hox11. *Genes Chromosomes Cancer*, **41**(4), 309–320.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**(6937), 241–254.

Killcoyne, S., Carter, G. W., Smith, J., and Boyle, J. (2009). Cytoscape: a community-based framework for network modeling. *Methods Mol Biol*, **563**, 219–239.

Kim, D., Xu, M., Nie, L., Peng, X.-C., Jimi, E., Voll, R. E., Nguyen, T., Ghosh, S., and Sun, X.-H. (2002). Helix-loop-helix proteins regulate pre-tcr and tcr signaling through modulation of rel/nf-kappab activities. *Immunity*, **16**(1), 9–21.

Moses, A. M., Chiang, D. Y., Pollard, D. A., Iyer, V. N., and Eisen, M. B. (2004). MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, **5**(12), R98.

Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D., and Eisen, M. B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*, **2**(10), e130.

Palomero, T., Odom, D. T., O'Neil, J., Ferrando, A. A., Margolin, A., Neuberg, D. S., Winter, S. S., Larson, R. S., Li, W., Liu, X. S., Young, R. A., and Look, A. T. (2006). Transcriptional regulatory networks downstream of tal1/scl in t-cell acute lymphoblastic leukemia. *Blood*, **108**(3), 986–992.

Pierce, J. W., Jamieson, C. A., Ross, J. L., and Sen, R. (1995). Activation of il-2 receptor alpha-chain gene by individual members of the rel oncogene family in association with serum response factor. *J Immunol*, **155**(4), 1972–1980.

Schotte, R., Rissoan, M.-C., Bendriss-Vermare, N., Bridon, J.-M., Duhen, T., Weijer, K., Brière, F., and Spits, H. (2003). The transcription factor spi-b is expressed in plasmacytoid dc precursors and inhibits t-, b-, and nk-cell development. *Blood*, **101**(3), 1015–1023.

Sinha, S., Adler, A. S., Field, Y., Chang, H. Y., and Segal, E. (2008). Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res*, **18**(3), 477–488.

Tone, Y., Kojima, Y., Furuuchi, K., Brady, M., Yashiro-Ohtani, Y., Tykocinski, M. L., and Tone, M. (2007). Ox40 gene expression is up-regulated by chromatin remodeling in its promoter region containing sp1/sp3, yy1, and nf-kappa b binding sites. *J Immunol*, **179**(3), 1760–1767.

Voronova, A. F. and Lee, F. (1994). The e2a and tal-1 helix-loop-helix proteins associate in vivo and are modulated by id proteins during interleukin 6-induced myeloid differentiation. *Proc Natl Acad Sci U S A*, **91**(13), 5952–5956.