**ChIP-chip peak detection**

NimbleGen RefSeq hg18 promoter tiling arrays (design ID 4226, design name 2006-07-18_HG18_RefSeq_promoter) were used for ChIP-chip analysis. Data normalization and peak detection was performed with the MA2C algorithm, the 4 U266 and 2 Daudi replicates were normalized using the robust method with C = 2 (1). Peaks were called using the \( p \) value method using a bandwidth of 500 and a \( p \) value cut-off of 0.001, with minimum probes set to 5 and max gap to 250. Data from the negative control Daudi cell line provided experimental assessment of false positive detection rates.

Predicted peak regions were linked to the nearest transcriptional start site (TSS) for RefSeq genes in the hg18 alignment (NCBI Build 36.1). TSSs within 3 kb in either orientation were accepted, unless an intervening TSS in the same orientation was present. For bi-directional promoters the nearest TSS was chosen as the primary association.

**Motif analysis**

**Denovo detection**

Two independent algorithms, Weeder (version 1.3 downloaded from http://159.149.109.9/modtools/downloads/weeder.html) and the web-based tool oligo-analysis (http://rsat.scmbb.ulb.ac.be/rsat/) were used to detect frequently occurring motifs in the set of peak regions (2, 3). They were chosen based on their performance compared to other methods in benchmark tests (4). Weeder is a consensus based method that enumerates all the oligos up to a maximum length and collects their occurrences (with substitutions) from input sequences. Oligo-analysis uses a set of algorithms that detect statistically significant motifs by counting the number of occurrences of each oligomer and comparing these with expectation. For both methods, we considered both strands of the input sequences and allowed a motif to appear more than once in sequence, otherwise we used default parameters. Motif lengths of 6, 8, 10 and 12 bases were searched for by Weeder. We restricted
the Oligo-analysis to search for motif lengths of 8 bases, the maximum length allowed by the web tool.

Sequence logos representing the position-weighted matrices generated by Weeder were drawn by the EnoLOGOS web tool using relative entropy as the output parameter (5).

**Random peak regions**

To provide a control for the *de novo* motif discovery techniques, a set of random sequences were generated. 500 genes were picked at random and a portion of sequence upstream of each TSS was selected. The length of each sequence portion corresponded to the mean length of MA2C predicted peaks, while the position of the sequence portion corresponded to the distribution of the mean distance from the centre of MA2C peak sequences to the closest associated gene. In this way, underlying sequence structure was retained, as opposed to sequences generated by random permutations of A, G, C and T. 5 independent sets of random sequences were generated by this processes and then submitted to the Weeder and oligo-analysis algorithms using the same parameters as before.

**STAMP-analysis**

The STAMP webserver ([http://www.benoslab.pitt.edu/stamp/](http://www.benoslab.pitt.edu/stamp/)) was used to compare the *de novo* generated matrices both against each other (pairwise comparison) and against either the TRANSFAC database (version 11.3) or the *in vitro* matrix derived for BLIMP1 (discussed below) (6, 7). For all analyses the default settings were used: Pearson Correlation Coefficient, Ungapped Smith-Waterman alignment, Iterative Refinement and UPGMA (distance based tree construction). The tree construction, via iterative refinement, produces an alignment of all of the input matrices along with a familial matrix that is the average profile of the multiple alignment.

**Position weight matrix analysis**

Using position weight matrices (PWM), in TRANSFAC-format, and the RSAT matrix-scan program we scanned both the MA2C predicted peak regions (286 peaks, average length 1352 bp) and every TSS in the hg18 reference genome (NCBI Build 36.1) from -2200 to + 500 bp (if multiple TSS overlapped they were joined into a larger region; 20,376 regions average length 2,797 bp) (6, 8). Regions were scanned
with all of the Weeder generated matrices (28; 14 all-occurrences and 14 best-occurrences) and the \textit{in vitro} weight matrix for BLIMP1 (\textit{in vitro} matrix). The BLIMP1 matrix used was derived from the high and medium affinity matrix defined by Kuo & Calame (9).

To eliminate zero counts, we assigned pseudo-counts in two different ways. For the \textit{in vitro} matrix we assigned counts based on prior knowledge of known BLIMP1 binding sites not included in the alignment. These pseudo-counts were equally distributed at positions 1-4 and 9-11 (each count from the original alignment was incremented by 0.25) but heavily biased to GAAA at positions 5-8 (set to \(0.999997/1*10^{-6}\); no variation has been seen at these positions in previously defined BLIMP1 binding sites). Thus the presence of GAAA becomes a pre-requisite for a high scoring motif. For the Weeder matrices we normalized the counts per base to 1000 and then incremented each count by 0.25. The original matrices are given in Supplementary Table2.

**Bootstrapping**

The RSAT matrix-scan program derives a \(p\) value for each PWM match giving an estimate of the likelihood of seeing a match by chance alone. Thus for a \(p\) value = 0.001 you would expect one match every kilobase. However, these estimates are not entirely accurate (especially at lower significances) and thus to gauge the enrichment and corresponding statistical significance of the different position weight matrices at different matrix-scan \(p\) values a bootstrapping approach was used. For each MA2C predicted peak a random TSS was selected from the total Nimblegen hg18 promoter tiling array set. The random peak was created to match the MA2C peak in both length and position relative to TSS. This process was carried out for all MA2C peaks, generating a random set of peaks that matches the MA2C peaks exactly for DNA length and very closely for relative position. To speed up bootstrapping the whole set of TSS from Nimblegen hg18 were pre-scanned with each PWM from -2200 to +500. If a MA2C peak lies outside of this range or is larger than 2700 bp then a peak is created that is as close as possible to the MA2C peak. As 2 of the MA2C peaks are larger than 2700 bp there is a discrepancy between the total MA2C peak’s DNA length and that of the total random peaks. This was dealt
with by randomly selecting MA2C peak TSS positions and creating new peaks until the total DNA lengths match exactly. At each of $10^6$ bootstraps the number of matrix-scan matches less than or equal to a given threshold were counted for both the MA2C and random peaks.

For each matrix a set of thresholds was generated by ranking all matches by $p$ value, and then determining a $p$ value threshold that identifies a given % of promoters that have at least one match with a $p$ value $\leq$ to that threshold. This procedure generates a set of normalized thresholds that allows for a direct comparison between the different matrices.

The bootstrapping results revealed that in almost all cases the Weeder best-occurrences matrices gave higher levels of enrichment than the Weeder all-occurrences matrices. Thus for all downstream analysis we used the Weeder best-occurrence matrices, from here on termed WWM (Weeder Weight Matrix).

Assessing overlap between different weight matrices

We assessed the overlap of the four WWMs corresponding to GTG or GNG based motifs with the longest 5’ or 3’ extensions (WWM9, WWM10, WWM11 and WWM13), in addition the matches from these 4 WWMs were combined and compared with the \textit{in vitro} matrix. If two matrices have a match for a promoter, regardless of the position of those matches, then it is deemed to be a shared promoter/gene.

Analysis of CpG content

The CpG content of all hg18 Refseq promoters (-2200 -> + 500 bp) and all 286 MA2C peak regions was calculated. This was then normalized to GC content:

$$\text{Obs/Exp CpG} = \frac{\text{Number of CpG}}{\text{Number of C} \times \text{Number of G}} \times N,$$

where $N$ is the total number of nucleotides in the sequence (10).

Selection of peaks with overlapping BLIMP1/IRF sites

Analysis of the overlap between the genes with significant Weeder Weight Matrix Quartet (WWM-Q; WWM9, WWM10, WWM11 and WWM13 combined) and \textit{in vitro}
matrix matches revealed that if the most significant 15% matches were considered then the WWM-Q captured all of the genes found by the \textit{in vitro} matrix along with an additional 43%. This threshold was then chosen to represent “high-confidence” BLIMP1 binding sites and used to analyse the overlap between BLIMP1 and IRF sites. For each WWM-Q match within the “high-confidence” set the match position was transformed to a reference position (relative to the \textit{in vitro} matrix). Thus in the forward direction WWM10 (AAGTGAAAGT) and the \textit{in vitro} matrix (AAGTGAAAGTG) both have the same offset of 0 and thus the WWM10 position in not altered. In contrast, the WWM13 (GTGAAAGTGA) has an offset of 2, and thus the position is modified by -2 bp. Following this process the DNA sequence at the transformed position is searched for the IRF core consensus (AANNGAAA). Any gene containing at least one such site was deemed to have an overlapping BLIMP1/IRF site.

\textbf{Analysis of conservation and occurrence of IRF sites in peak regions}

Data on primate (human, chimp, gorilla, orangutan, rhesus, marmoset, tarsier, mouse lemur and bushbaby) sequence conservation was downloaded from the University of California, Santa Cruz (UCSC) genome browser for the PhyloP and PhastCons methods (11, 12). For the WWM-Q and the \textit{in vitro} matrix a set of non-redundant PWM matches was generated. For all matrix matches that overlap only the most significant match was maintained and the rest were discarded, giving a list of the \textit{best matches per position} (BMPP). For each of matches in the BMPP list the conservation was calculated as the sum of the individual bp conservation values normalised by the match sequence length. The BMPP list was then ranked by these normalised conservation values, from highly conserved sites to those with low conservation (or accelerated evolution in the case of PhyloP). A comparison was then made between the fraction of BLIMP1 sites with IRF overlap versus those without IRF overlap in the top and bottom of this ranked list for a given percentage (e.g. top 5% of the list versus the bottom 5%). A Fischer’s exact test was then used to examine the significance of association of overlapping BLIMP1/IRFs in the top/bottom portion of the ranked list.
**Expression Analysis**

**Expression array data analysis**

Gene expression was analysed using Nimblegen expression arrays (HG18_60mer_expr, design for human [hg18; NCBI Build 36]). These chips contain 4 individual arrays (4-plex) each with 72,000 60-mer probes with 3 probes per gene, thus providing data for 24,000 genes. In total 16 samples were analyzed on 2 separate chips, 8 BLIMP1 knockdown and 8 controls (randomized siRNA). The provided raw data pair files were converted to Nimblegen XYS format using a perl script. These files were then read into the R statistics package using the oligo R package (16). Assessment of images of the raw calls revealed little sign of spatial defects across the 16 arrays. However, some of the arrays appeared to have abnormally high or low overall signals. The raw data was further inspected using histograms and all-by-all comparisons of correlation. From this analysis 6 arrays (3 BLIMP1 knockdown, 3 controls) were deemed to have abnormal distributions of signals and were therefore removed from further analysis. The remaining 10 arrays were normalized using the Robust Multichip Average (RMA) method of the Oligo package (17). Finally, a linear model was fit to the data using the R Limma package and the significance of the differential expression was gauged using the Limma empirical Bayes statistics module (18). 85 probe triplets, accounting for 82 unique genes, were shown to have differential expression with an adjusted probability (Benjamini & Hochberg method) of < 0.1 of which 79 show greater than 2-fold change in expression.

**Assessing expression changes upon knockdown**

To test if there was a significant enrichment of the MA2C genes (target genes) amongst the genes showing increase/decreased expression upon BLIMP1 knockdown a comparison with the hypergeometric distribution was made. Firstly, for any gene that has probe triplets for alternative isoforms (e.g. UBE2L6 has 3 probes for both NM_004223 and NM_198183) only the isoform with the highest adjusted probability was kept and the rest were discarded, then the expression data was ordered from +ve to –ve fold change. Significance was then gauged by treating genes with increased/decreased expression (between BLIMP1 knockdown and
control) as separate populations and viewing a given list of target genes as the drawn sample for a hypergeometric analysis. The resulting probability is the likelihood of randomly drawing more genes with increased/decreased expression than observed, given a draw size equal to the number of target genes.

References

1. Song, J.S., Johnson, W.E., Zhu, X., Zhang, X., Li, W., Manrai, A.K., Liu, J.S., Chen, R. and Liu, X.S. (2007) Model-based analysis of two-color arrays (MA2C). *Genome biology*, 8, R178, 10.1186/gb-2007-8-8-r178. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17727723.

2. Defrance, M., Janky, R., Sand, O. and van Helden, J. (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nature protocols*, 3, 1589-603, 10.1038/nprot.2008.98. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18802440.

3. Pavesi, G., Mereghetti, P., Mauri, G. and Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research*, 32, W199-203, 10.1093/nar/gkh465. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15215380.

4. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23, 137-44, 10.1038/nbt1053. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15637633.

5. Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D. and Benos, P.V. (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic acids research*, 33, W389-92, 10.1093/nar/gki439. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15980495.

6. Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31, 374-8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12520026.

7. Mahony, S. and Benos, P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic acids research*, 35, W253-8, 10.1093/nar/gkm272. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17478497.

8. Turatsinze, J., Thomas-Chollier, M., Defrance, M. and van Helden, J. (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature protocols*, 3, 1578-88, 10.1038/nprot.2008.97. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18802439.
9. Kuo, T.C. and Calame, K.L. (2004) B lymphocyte-induced maturation protein (Blimp)-1, IFN regulatory factor (IRF)-1, and IRF-2 can bind to the same regulatory sites. *Journal of immunology (Baltimore, Md. : 1950)*, 173, 5556-63. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15494505.

10. Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *Journal of molecular biology*, 196, 261-82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3656447.

11. Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.a., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., et al. (2010) The UCSC Genome Browser database: update 2010. *Nucleic acids research*, 38, D613-9, 10.1093/nar/gkp939. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19906737.

12. Siepel, A., Pollard, K.S. and Haussler, D. New Methods for Detecting Lineage-Specific Selection Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.84.6412.

13. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25, 25-9, 10.1038/75556. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10802651.

14. Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, 21, 3448-9, 10.1093/bioinformatics/bti551. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15972284.

15. Killcoyne, S., Carter, G.W., Smith, J. and Boyle, J. (2009) Cytoscape: a community-based framework for network modeling. *Methods in molecular biology (Clifton, N.J.)*, 563, 219-39, 10.1007/978-1-60761-175-2_12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19597788.

16. Carvalho, B., Bengtsson, H., Speed, T.P. and Irizarry, R.a. (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics (Oxford, England)*, 8, 485-99, 10.1093/biostatistics/kxl042. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17189563.

17. Irizarry, R.a., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4, 249-64, 10.1093/biostatistics/4.2.249. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12925520.

18. Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3, Article3, 10.2202/1544-6115.1027. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16646809.
Doody et al

Supplementary Figure 1. ChIP sample evaluation pre- and post- T7 amplification. Representative real-time PCR data from BLIMP1 and control IgG ChIP samples prepared from U266 cells and evaluated at CIITA promoter-IV before and after T7 amplification. For T7 amplified samples PCR was performed on samples diluted 500-fold. Shown are the average absolute Ct value differences between duplicate PCRs, with errors bars showing deviation between samples. Also shown are the results at a negative control promoter fragment, an upstream region of the KRT10 promoter, after T7 amplification. This primer pair falls at the far margin of the peak region subsequently identified in ChIP-chip data and the fold enrichment at the corresponding tile in the primary data is log2 ratio=0.990.

Supplementary Figure 2. Additional motif analysis. (A) Weeder derived matrices were compared using the STAMP program. The familial motif tree drawn by STAMP using default settings for the 14 Best Occurrence (BO) versions of Weeder derived matrices is shown. (B) Logos representing all 14 BO Weeder matrices and the high-affinity in vitro weight matrix derived for BLIMP1 from Kuo & Calame (2004). Beneath each matrix is the best PWM match identified by STAMP in the TRANSFAC v11.3 database along with the corresponding E-value. Beneath this in the dark grey box is the E-value given by STAMP for matches to the High-affinity BLIMP1 matrix, which is not contained in TRANSFAC v11.3. (C) Logos of the in vitro matrix, WWMs and the IRF cores consensus (dark boxes). For comparison: the IRF core consensus and BLIMP1 and IRF PWMs from TRANSFAC (light grey boxes).

Supplementary Figure 3. Observed vs expected occurrences of WWM-9 in wild type and “in silico mutated” forms. The distribution of motif matches in human RefSeq promoters was determined using the RSAT matrix-scan program in $10^6$ randomly selected sets of promoter regions matching BLIMP1 bound regions for length, position and number. (A) Graphs of distributions of motif matches equal to or better than four motif quality thresholds ($pVal\leq0.001$, $\leq0.0001$, $\leq0.00001$, $\leq0.000001$)
for WWM-9 (the representative logo is shown at the top). The observed number of matches in the BLIMP1 bound peak regions is shown as a red-line. (B) The same analysis was performed with a “mutated” matrix in which the values for G and C were swapped at critical positions, in this case for WWM-9Mut, positions 3 and 9 as shown in the logo and in Supplementary Table 2. The observed frequency of motif matches to this matrix, WWM-9Mut, in occupied promoter regions (red line) is markedly different from the original WWM-9 in (A).

Supplementary Figure 4. Conserved BLIMP1/IRF sites have better quality motif matches. A list of all non-redundant BLIMP1 motif predictions was generated for the WWM-Q and in vitro matrix. For each site a conservation score was generated from the sum of PhyloP scores at each position of the match, using either PhyloP scores for primate (upper panels) or vertebrate (lower panels) conservation. The resulting ranked list was divided into corresponding fractions of top most (black line) and bottom least (red line) conserved motif matches (fractions from left to right: 10%, 20%, 40%). Within each of these fractions motif matches were divided into IRF overlap or non-IRF overlap and the corresponding distributions of p values were plotted.

Supplementary Figure 5. IRF1 can occupy BLIMP1 target promoters with overlapping BLIMP1/IRF sites. Real-time PCR quantification of IRF1 promoter occupancy in duplicate ChIP samples from U266 cells. A set of BLIMP1 target promoters with overlapping BLIMP/IRF binding sites was evaluated. Data is displayed as mean fold enrichment ± SD relative to control rabbit IgG. Using 3-fold enrichment as cut-off 8 of 9 promoters tested are bound by IRF1.

Supplementary Figure 6. Motif type per se is not predictive of a change in gene expression. Global gene expression changes in siRNA treated samples were evaluated on NimbleGen micro-arrays. The enrichment of three groups of genes amongst all genes showing an increase in expression above fold-change cut-offs from 1.4-fold to 2-fold was evaluated using a hypergeometric test. The groups of genes considered were: group (i) genes as in Figure 7 C, i.e. genes with BLIMP1 occupied promoters containing a top 15% WWM-Q match with IRF overlap (black fill), all RefSeq genes associated with a top 15% WWM-Q match with IRF overlap
regardless of BLIMP1 occupancy (light diagonal stripes), or to this set after removal of group (i) genes, giving the set of genes linked to unoccupied BLIMP1/IRF sites (dark diagonal stripes). The observed enrichment for each group amongst genes changing expression is shown in the upward bars, the significance of this enrichment is shown as the log10 of the $p$ value in downward bars. The dotted line represents $p \leq 0.05$. The minimal significance observed for the set of all RefSeq genes with BLIMP1/IRF matches (light diagonal) is eliminated once occupied genes are removed from the set (dark diagonal).