Immunity, Volume *31*

Supplemental Data

Analysis of Interleukin-21-Induced *Prdm1* Gene Regulation Reveals Functional Cooperation of STAT3 and IRF4 Transcription Factors

Hyokjoon Kwon, Danielle Thierry-Mieg, Jean Thierry-Mieg, Hyoung-Pyo Kim, Jangsuk Oh, Chainarong Tunyaplin, Sebastian Carotta, Colleen E. Donovan, Matthew L. Goldman, Prafullakumar Tailor, Keiko Ozato, David E. Levy, Stephen L. Nutt, Kathryn Calame, and Warren J. Leonard

Supplemental Experimental Procedures

Mice and Cell Culture. C57BL/6 mice were from the Jackson Laboratory. *Stat3*-deficient T and B cells were generated by crossing *Stat3*-floxed mice with p56$^{lck\text{-cre}}$, CD4-cre, and CD19-cre transgenic mice (Lee et al., 2002). *Irf4*$^{-/-}$ and *Irf8*$^{-/-}$ mice (Holtschke et al., 1996; Mittrucker et al., 1997), and *PU.1* deficient B cells (Polli et al., 2005) were described. Mice were analyzed at 4–9 weeks of age. Animal experiments used protocols approved by the NHLBI Animal Use and Care Committee and followed NIH guidelines. M12, CH12, and Bcl-1 B cell lymphoma lines and primary splenic B and T cells were cultured at 37°C in RPMI 1640 medium containing 10% fetal bovine serum (FBS), 2 mM L-glutamine, 100 U/ml penicillin G, and 100 μg/ml streptomycin, and 55 μM β-mercaptoethanol. NFS201 and NFS202 cells were cultured as described (Lee et al., 2006).

Quantitative Real-time PCR analysis. Total RNA was extracted using TRIzol (Invitrogen, Carlsbad, CA). First-strand cDNA was made from 2 μg of total RNA using random hexamers and Omniscript reverse transcriptase (QIAGEN, Valencia, CA) per the manufacturer’s protocol.

Quantification of specific mRNAs and control 18S rRNA was performed by quantitative real-time PCR using the 7900H sequence detection system (Applied Biosystems, Foster City, CA). cDNAs were amplified using TaqMan universal PCR master mix (Applied Biosystems). Primers and probes used to detect murine *Prdm1* mRNA were: forward primer 5'-ACACAGGAGAGAAGCCACATGA-3', reverse primer 5'-GGTGGGTCTTGAGATTGCTTGT-3', and probe 5'-[6-FAM] TGCCAGGTCTGCCACAAGAGATTTAGCA[TAMRA-6-FAM]-3' or from ABI. *Bcl3*, *Socs3*, *Tha1*, *Pim1*, and 18S rRNA primers and probes were from ABI.

Enhancer Constructs. PCR fragments containing mouse *Prdm1* regions were generated using *PfuUltra* Taq polymerase (Stratagene, La Jolla, CA), and inserted 5’ of the *Prdm1* promoter between the *Kpn* I and *Sac* I sites in the polylinker. Site-directed mutagenesis or deletions of these regions in pGL4-Basic were made using QuikChange (Stratagene) and verified by sequencing.

Electrophoresis Mobility Shift Assays (EMSA). Nuclear extracts were prepared from NFS201 cells untreated or treated with 50 ng/ml IL-21 for 7 h at 37 °C. EMSAs were performed as described (John et al., 1996) using 5% polyacrylamide gels (29:1 acrylamide:bis) in 0.5 × Tris-borate-EDTA buffer. For supershifting assays, nuclear extracts were pre-incubated for 10 min with antibodies.

Chromatin Immunoprecipitation assays.

Chromatin immunoprecipitation (ChIP) assay were done as described (Moreno et al., 1999). Splenic B and T cells were preactivated, and not stimulated or stimulated with 100 ng/ml of mouse IL-21 for 30 min at 37 °C, followed by cross-linking with formaldehyde. Nuclear lysates were immunoprecipitated with antibodies to STAT3, IRF4, and IRF8 or with control IgG. After treatment with proteinase K and reversal of cross-links, selected DNA sequences were assessed by quantitative real time PCR. Primers and TaqMan probes were as follows: mouse *Prdm1* enhancer: forward primer, 5'-GCAGCCCGAACCCCTTAA-3', reverse primer, 5'-CTGGAGGCAATCACAACGAA-3', and probe, 5'-(6-FAM)-CCACTGCTGCACTGGGCTCGG-(TAMRA-6-FAM)-3', mouse *Actb*, third intron of reference mRNA NM_007393: forward primer 5’-CAGAAAGCCACAAGAAACACTCA-3’, reverse primer 5’-ACTCCCAGCACACTGAACTTAGC-3’ and probe 5’-[FAM]-AGATCTGAGACATGCAAGGAGTGCAAGAACA-[TAMRA]-3’.

Statistical analysis of microarray.

Primary expression analysis was conducted with the Affymetrix GeneChip Operating System (GCOS), version 1.4 client software. Expression data were transformed using a variance stabilizing, quantile normalized function termed "S10" Comparative analysis between expression profiles for samples with or without treatments was carried out using MSCL Analyst's Toolbox developed for the JMP statistical software package (http://abs.cit.nih.gov/geneexpression.html; SAS Institute, Cary, NC). One-way analysis of variance (ANOVA) for time and treatment was used to derive *P* values for each probe set, and differentially regulated genes were selected using $P \leq 0.001$ and 1.5-fold cut off. False discovery rates (FDRs) were calculated and probes having less than 10% FDR were selected for further analysis. Fold changes in response to treatment were calculated as differences of mean S10 values for each treatment category. When multiple probe sets for a single gene were available, data were summarized by selecting the most extreme probe set fold-change. Hierarchical cluster analysis was computed using the Ward's method based on deviation of S10 expression values from the mean.

ChIP Seq data processing.

1. All the raw sequences were stacked, and the nucleotides at each position were counted (Figure S4A). The sequence tags were screened at low stringency against the linker or any other sequence readable in the ATGCN sequence profile.
2. They were then mapped using the AceView aligner to the mouse reference genome 37, allowing for up to 2 mismatches, single base insertions, deletions, or substitutions per tag (Figure S4B). Because the tags were only 25 bp long, 22 to 26% mapped in more than one position in the genome: only the tags having a unique best match genome wide were kept.
3. Each tag was replaced by a Gaussian of surface 1 and width σ constituting a "building block". This Gaussian convolution dampens the sampling fluctuations and introduces a controlled level of “fuzziness” (Figure S4C). The value for σ was optimized by visual inspection and depends on the depth of the experiment; we used σ =100 bases. The sum of the elementary Gaussians represents the tag density.
4. Each tag was shifted 3' by half the effective length of the sonication fragments. This length was estimated as the maximum of the genome wide correlation function between the densities of tags mapped on the plus and on the minus strand of the chromosomes (Figure

S4D). For an unknown reason, this value was significantly shorter than the length of the sonication products measured on the gel.

5. A ubiquitous low level of background noise, of the order of 1 tag per kb, was observed. Sharp concentrations of tags into peaks, interpreted as binding sites, were also seen. Peaks were identified as regions where the tag density exceeded a base threshold of one tag per 50 bases (Figure S5A).
6. The area of the peaks was measured, and the area of the local matching immunoglobulin control experiment was subtracted. The resulting area represents the number of tags attributable to direct or indirect binding of the specific protein, as the background immunoglobulin binding was removed. The width of the peak, measured at the level of the base threshold, indicates the spread and the complexity of the binding site. Small peaks had an average width of 250 bp. A finer resolution of the anatomy of high peaks (more than 50 tags) was obtained by narrowing the Gaussian representation of each tag to 50 or even 20 bases (as exemplified for *Prdm1* in Figure 6B).
7. To choose an area threshold distinguishing the signal from the noise, two histograms were plotted: the number of peaks with area n tags represented the signal, whereas the number of genomic segments, selected outside of the peaks and of length 250 bp that were hit by n tags represented the noise (Figure S5B). For each experiment, to our surprise, the two distributions were log-linear: in semi-log coordinates, the histograms fit two lines with different slopes. The intersection of the two lines defines the threshold for the specific experiment. Peaks with lesser area were obscured by the background noise and thus removed, whereas peaks with greater area were retained. The number of false positives was estimated as the area of the triangle above the threshold and below the steep background

noise distribution (Figure S5B). Interestingly, the observed log-linear distributions depart from the usually considered Poisson distributions: they are more spread, have a larger tail at high counts and a higher number of false negatives. If m is the average number of tags per peak, the fraction of peaks with no tags is p(0)= 1/(m+1), which is larger and a softer function of m than the Poisson value p(0)= e^{-m}.

8. Peaks were clustered into ‘sites’ by transitive contact: if the sequences of two peaks from different experiments overlapped, they were considered part of a single binding site whose width was the combined width.
9. If a site contained a peak from an immunoglobulin control track, peaks from the corresponding experiments with area less than four times the control peak area were removed. Rare control immunoglobulin peaks exceeded 100 tags, and those regions were removed from the analysis.
10. The set of DNA sequences associated to the binding sites was scrutinized, as the most frequent “words” identify protein binding motif candidates. A systematic search for frequent words containing between 5 and 11 letters, with or without gaps (up to 3 letter gaps), returned lists sorted by frequency, that were then leveled by the a priori probability of occurrence of the word, estimated as its frequency of occurrence in the genome. As we are studying transcription factor binding sites, we hypothesized that binding would target mainly double stranded DNA; hence word candidates were considered “interesting” only if they were palindromic or if the sense and antisense sequences were found at similarly high frequencies.
11. To evaluate the importance of the candidate words in the protein binding process, a histogram of the distance between the maximum of the peak and the central letter of the word

was plotted for each experiment and each candidate word. A word with highly skewed position distribution in the sum of all sites is most likely biologically meaningful to the interaction: centered histograms, where the word's favorite location is in the center of the peaks, at the maximal tag density, suggest direct involvement of the word in binding site recognition. A skewed out-of-center distribution is equally informative, as it might correspond to a secondary site for the tested protein or a primary site for a partner protein in the binding complex. Flat distributions suggest the word may be frequent but random in its position, making its relevance to binding questionable. The histogram for the exact GAS motif matched a Gaussian with $\sigma = 10$, centered at zero with one base accuracy (Figure S5C). This level of precision validates the protocol used for the analysis. Among the single letter variants of the GAS motif, TTCnnnTAA was the best centered.

12. STAT3 binding sites mainly coincide with IRF4 binding sites (Figure S6). However, the relationship between the positions of the STAT3 binding sites and the genes whose expression STAT3 might influence is not trivial. Each gene is potentially associated to many candidate sites lying within the gene, or upstream or downstream from the gene, up to hundreds of kilobases away. Conversely, each binding site may be associated to multiple genes in its vicinity. We selected the candidate genes in the following way: if a site lies within one or several genes, at least one of which has an NCBI Entrez GeneID, these genes are the unique candidates associated to the site. Otherwise, we consider as candidates the two nearest left and right genes with an NCBI Entrez GeneID, additionally in that interval, the nearest spliced gene supported by cDNAs in GenBank and annotated in AceView, and finally also the nearest single-exon AceView gene if closer than the nearest spliced gene.

13. Correlation to Affymetrix expression results was analyzed (Figure S7A and S7B), and we found that only 14% of the STAT3 binding sites map within 1 kb from the transcription start site of genes whose expression is modified by the IL-21 treatment. *Prdm1* is an example of a gene where a relevant site is located downstream of the regulated gene.

Keyhole Limpet Hemocyanin (KLH) Immunization. *Irf4*$^{-/-}$ and littermate control mice (8–10 wk old) were immunized with KLH (Nurieva et al., 2008), sacrificed 8 d later, and lymph nodes analyzed. The germinal center B cells were determined by staining with FITC-PNA (Vector Laboratories, CA) and APC-anti-B220 mAb (PharMingen). The Tfh cells were determined by staining with APC-anti-CD4, PE-anti-CXCR5, and biotinylated anti-ICOS mAb (PharMingen), followed by FITC-labeled streptavidin (PharMingen).

Table S1. Sequencing depth and mapping summary. Shown are data from two independent ChIP-Seq experiments.

Two combined replicas	IRF-4 bound before IL-21	IRF-4 bound after IL-21	Stat3 bound before IL-21	Stat3 bound after IL-21
Number of uniquely mapped tags	6,271,629	5,589,355	5,710,866	4,754,238
Threshold	9	9	15[2]	9
Expected false positives (%)	450 (3.1%)	190 (1.2%)	17 (NA)	187 (4.2%)
Number of tags in peaks	357,685	407,789	9,789	124,042
Number of peaks[1]	14,722	15,867	335	4,478
Average tags per peaks	24.3	25.7	29.0	27.7
Average peak width (bp)	410	409	407	389
kb of sequence bound genomewide	5,909	6,497	137	1,742

[1] The number of peaks is slightly larger than the number of sites (Figure 6A), because peaks that are sufficiently close to each other may become merged into a single site through transitive contact.

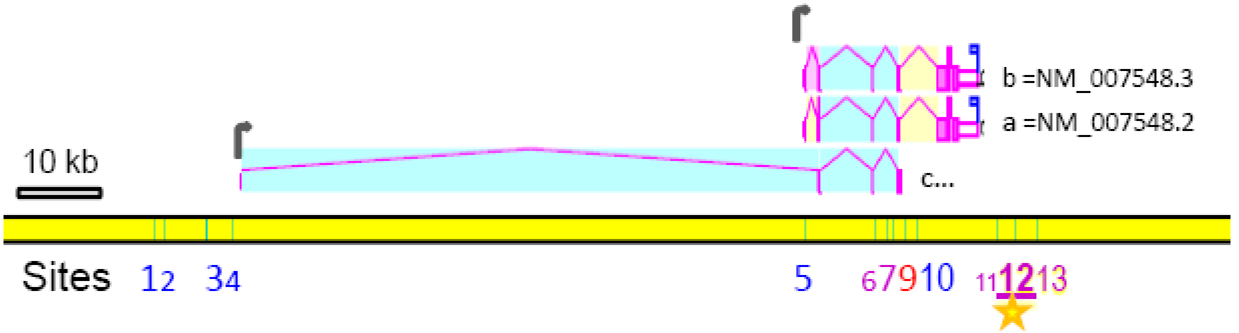
[2] This very high value corresponds to the fact that the noise dominates in this experiment, as there are essentially no STAT3 binding sites before IL-21 treatment.

Table S8. Genes associated to both STAT3 and IRF4 sites are more prone to being regulated by IL-21.

Best candidate site associated to the gene	Genes with gene ID tested by Affymetrix	Expression is sensitive to IL-21	% differentially expressed genes (± Wilson confidence interval)
Genes with at least one nearby site binding both STAT3 and IRF4	2,754	729	26.5 ± 1.6
Genes with a nearby site binding STAT3 and another binding IRF4, but none binding both proteins	426	86	20 ± 4
Genes with a nearby site binding STAT3 but none binding IRF4	193	28	15 ± 5
Genes with a nearby site binding IRF4, but none binding STAT3	7,655	933	12.2 ± 0.8
Genes lacking a nearby candidate binding site	6,801	431	6.3 ± 0.6
Total	17,829	2,207	12.4 ± 0.5

Entrez genes tested on the Affymetrix array and whose expression depends on IL-21 (at the threshold of 1.5 fold) were partitioned according to the type of protein binding sites in their vicinity. The percentage of differentially expressed genes and its confidence interval show that IL-21 regulated genes are markedly enriched near STAT3 or IRF4 binding sites, but much more so in the vicinity of a composite site binding both proteins.

Table S11. Binding sites near the *Prdm1* gene, in CD4$^+$ T cells, before and after IL-21 induction.



Site num ber	Coordinate of composite site center (chr 10 build 37)	Position relative to *Prdm1* gene (in AceView)	STAT3 or IRF4 bound, before of after IL-21	Number of tags in the binding site	GAS or GAS-like motif in site
1	44259130	10,823 bp upstream of variant c promoter	IRF4 no IL21 IRF4 + IL-21	25 tags 25 tags	1 GAS TTC…GAA
2	44257890	9,581 bp upstream of variant c promoter	IRF4 no IL-21 IRF4 + IL-21	17 tags 17 tags	
3	44252510	4,022 bp upstream of variant c promoter	IRF4 no IL-21 IRF4 + IL-21	21 tags 24 tags	
4	44249395	1,088 bp upstream of variant c promoter	IRF4 no IL-21 IRF4 + IL-21	14 tags 18 tags	
	44248308	5’ end of gene (5’ of variant c Sep07)			
5	44178520	Promoter of a and b; 31 bp after start of transcription of a	IRF4 no IL-21 IRF4 + IL-21	**25 tags** 15 tags	
6	44169670	Intron 4 of variant a (intron 3 of variant b/c)	STAT3 + IL-21 IRF4 + IL-21	12 tags **10 tags**	1 GAS and 1 TTC…GCA, 137 bp apart
7	44168060	Intron 4 of variant a (intron 3 of variant b/c)	IRF4 no IL-21 STAT3 + IL-21 IRF4 + IL-21	16 tags 30 tags **34 tags**	1 GAS and 4 TTC…TAA, successively separated by 61, 109, 79, and 248 bp
8	44167370	Intron 4 of variant a (intron 3 of variant b/c)	IRF4 + IL-21	**10 tags**	
9	44166018	Intron 5 of variant a (intron 4 of variant b)	STAT3 + IL-21 IRF4 + IL-21	61 tags **12 tags**	1 GAS TTC…GAA
10	44164433	Intron 5 of variant a (intron 4 of variant b)	IRF4 no IL-21 STAT3 + IL-21	58 tags 15 tags	

			IRF4 + IL-21	**79 bp**	
	44156975	3’ end of gene (3’ end of variants a/b Sep07)			
11	44154623	2.3 kb 3’ to the gene	STAT3 + IL-21 IRF4 + IL-21	16 tags **10 tags**	1 GAS TTC…GAA
12*	44152250	4.7 kb 3’ to the gene	IRF4 no IL-21 STAT3 + IL-21 IRF4 + IL-21	64 tags 70 tags **137 tags**	1 TTC…TAA
13	44149503	11.9 kb 3’ to the gene	IRF4 no IL-21 STAT3 + IL-21 IRF4 + IL-21	21 tags 66 tags 30 tags	2 TTC…TAA (360 bp apart)

The top of the diagram shows the structure of the mouse *Prdm1* gene, copied from a unique NCBI transcriptome database called AceView that integrates into genes and alternative mRNA variants all cDNA sequences from the public repositories (GenBank, dbEST or Trace Archive at NCBI) (www.aceview.org i.e. www.ncbi.nlm.nih.gov/IEB/Research/Acembly). In the current version (dated Sep07), the mouse *Prdm1* gene includes three alternative transcript variants, a, b, and c Sep07, defined by the sequences of 49 independent cDNA clones in all public sequence databases. The reference mRNA sequence, NM_007548.2 (dated November 17, 2006) corresponds to AceView variant a (Sep07), which is supported by U08185, a cDNA isolated by Turner et al. from B lymphocytes (BCL1 cells induced by IL-2 + IL-5) (Turner et al., 1994). The predominant form b (Sep07) is mainly found in eye and ovary; it differs from variant a by skipping the second (79 bp) exon and corresponds to the current version of RefSeq, NM_007548.3 (dated April 4, 2008). A third variant, c Sep07, shows the existence of a second promoter, 70 kb upstream from the main promoter, which is supported by accession AK077622 from an 8-day embryo (from Riken cDNA clone clone:5730478J08); its sequence is currently only known for the 5’ end.

To summarize, the *Prdm1* mouse gene has two promoters: the most 5’ promoter is active in embryos and the main promoter 70 kb downstream is active in eye, spleen, thymus, ovaries and skin. The second exon of variant a is an alternative 79 bp cassette skipped in B cells. As a result, the three *Prdm1*–encoded protein isoforms expressed in either eyes and ovary, B cells, or embryos have different N-termini.

Analysis of the ChIP-Seq profiles (Figure 6C) identified 13 STAT3 or IRF4 binding sites in the vicinity of *Prdm1*, numbered 1 to 13 from 5' to 3' along the gene. Binding site 12 is denoted as 12*, to indicate that it corresponds to the IL-21 response element discussed in detail in this study. Binding sites are represented by thin lines in the yellow bar below the gene; their characteristics are detailed in the table. Protein binding peaks with more than 30 tags are highlighted in yellow. For each composite binding site numbered in column 1, the table shows

- the coordinate on chromosome 10 (column 2, on mouse genome build 37) of the center of the composite binding site generated by merging eventual overlapping peaks from the four types of experiments, STAT3 or IRF4, with or without IL-21, where each experiment combines two independent replicas and subtracts their two related IgG controls.

- The position of the site relative to the *Prdm1* gene (column 3, as depicted in AceView Sep07): the transcribed part of the gene *Prdm1* spans 91.33 kb on chromosome 10, from position 44248308 to position 44156975 in NCBI mouse genome build 37. Four sites (sites 1 to 4) lie upstream of the first promoter, one (site 5) lies inside the second promoter, five more (sites 6 to 10) are within the transcribed gene, and three (sites 11 to 13) are downstream of the 3' end. Note that due to alternative splicing and the complexity of the transcription pattern, describing positions relative to exon or intron numbers is imprecise and can be misleading: for instance, intron 4 of variant a is intron 3 of variants b and c.

- The peak area, which measures the strength of the binding and reflects the number of tags in each binding peak, is provided in columns 4 and 5. This number is computed after an elaborate treatment of the Solexa sequence data, including tag shifting, thresholding, Gaussian smoothing, and substracting control IgG tags.
- The number of motifs of the GAS type that we demonstrated (Figure S3) are relevant to STAT3/DNA binding (i.e. TTC…GAA, TTC…TAA/TTA…GAA or TTC…GCA/TGC…GAA) is reported in column 6. When multiple motifs are found within the same peak, the inter-motif distance in bp is indicated.

The reporter assays and other experiments described in Figures 2-5 tested sites 5 to 12*. Only site 12*, 4.7 kb downstream of the gene and which contains the 212 bp IL-21 response element, appears functionally critical (Figures 2-5); ChIP-Seq results for this very strong binding site are shown in detail in Figure 6B.

Before IL-21, no binding of STAT3 is observed, consistent with the fact that STAT3 is then expected to be unphosphorylated and mainly cytosolic, but seven STAT3 binding sites appear after IL-21 treatment (sites 6, 7, 9 to 13, red lines in the table). Those binding sites all map in the distal part of the gene, lying from 13 kb upstream to 12 kb downstream of the 3’ end of the gene. Remarkably, all 7 STAT3 IL-21 induced sites overlap IL-21 induced IRF4 binding sites, and in all cases IRF4 binding is increased after IL-21 treatment. Intronic sites 7 and 10 as well as sites 12* and 13, downstream of the gene, also overlap pretreatment IRF4 sites. STAT3 sites 6, 7 and 9 contain a canonical GAS motif, whereas sites 7, 12* and 13 contain respectively 4, 1 and 2 TTC…TAA variants, shown to be critical to STAT3 binding in site 12* (Figure 3).

All 13 sites bind IRF4 after IL-21 treatment; nine of those were bound to IRF4 even before IL-21 induction. Notably, in un-stimulated or IL-21-stimulated CD4$^+$ T cells, there is clear binding of

IRF4 near the two promoters of the gene, but none of the promoter sites bind STAT3. Four sites (1 to 4) map within 11 kb upstream of the 5’-most promoter, and binding there is insensitive to IL-21. Site 5 binds right at the transcription start of the second and main promoter, used in particular in lymphoid tissues; this site is the only one where IL-21 induction leads to a weakening of IRF4 binding.

Figure S1. IL-21 induces *Prdm1* in splenic B cells. (A) Pre-activated splenic B cells were treated with 100 ng/ml IL-21, and *Prdm1* mRNA levels determined at indicated times. (B) IL-21-induced *Prdm1* expression is faster but weaker/less sustained than with LPS. Pre-activated splenic B cells were treated with 50 ng/ml IL-4, 100 ng/ml IL-21, or 3 mg/ml LPS for various time points, and *Prdm1* mRNA levels (means ± SEM of 3 independent experiments, total combined samples ≥ 4) were determined by quantitative RT-PCR.

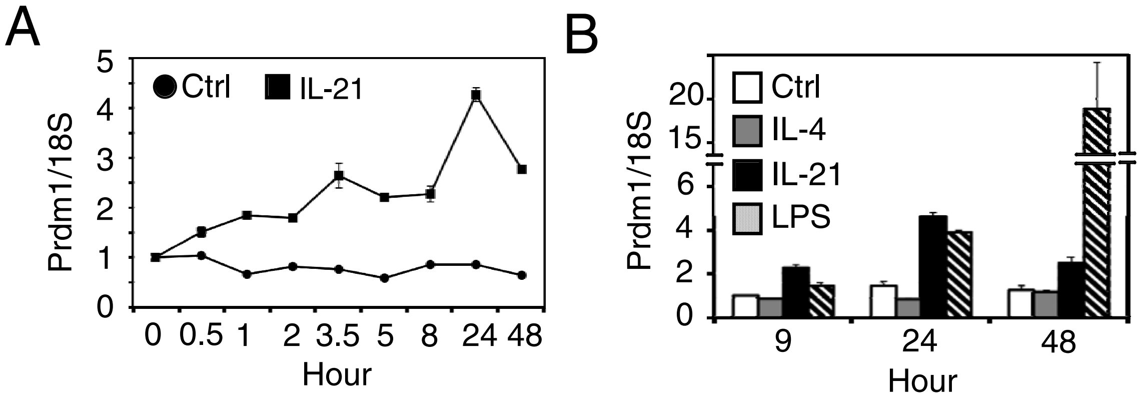


Figure S2. Principal component assay (PCA) of IL-21-induced gene expression in CD4$^+$ T cells. PCA of five independent samples of CD4$^+$ T cells stimulated with IL-21 for Control (Ctrl, blue), 1 (green), 6 (purple), or 24 h (red).

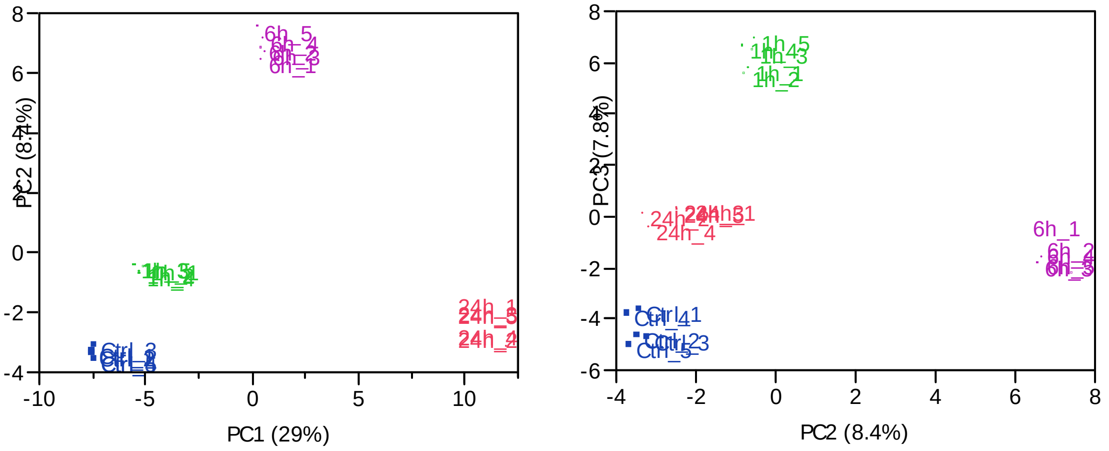


Figure S3. IL-21-induced gene expression in CD4$^+$ T cells. Clustering Analysis of Affymetrix gene expression. Five independent samples of CD4$^+$ T cells stimulated with IL-21 for 0, 1, 6, or 24 h. *P* ≤ 0.001, 1.5-fold cutoff.

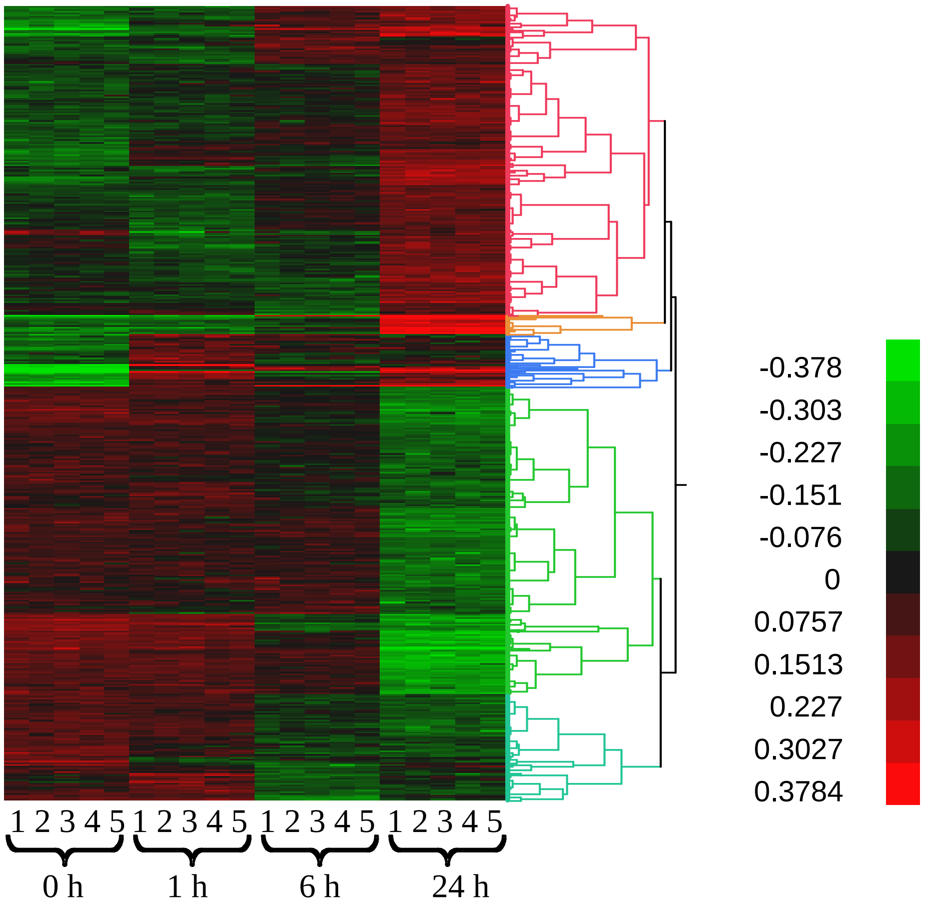


Figure S4. ChIP Seq analysis. (A) Plotting the base composition of the sequence tags reveals the primer sequence GATCGGAAGAGCTCGTATGCCGTCT. The tags matching this sequence with up to 6 mismatches are exhaustively searched and excluded. (B) The pattern of mismatches depends on the Illumina/Solexa run (experiment) more than on the particular sample; insertions and deletions are rare, and substitutions are nonrandom. The quality of the sequencing data has continued to improve rapidly, and recent Illumina runs show fewer mismatches. (C) Impact of Gaussian smoothing: Example of STAT3 (red) and the corresponding mouse immunoglobulin control (black) binding in the *Hivep*2 gene area. The area shown is 2 kb long. The y axis indicates the number of tags inside a monomodal peak. Each mapped Solexa tag is spread as a Gaussian with tunable width, indicated on the left. As σ increases, each tag contributes density to points further away. The local number of tags dictates which σ should be used to get the best resolution. By default, we use 100 basepairs, but in this example 50 bp is optimal to see the substructure of the binding site, which appears to contain distinct sub-regions where the protein binds. (D) Correlation between tag densities on the two strands yields the effective length of the sonication fragment. Mapped tags are translated downstream by half the length measured at the maximum of these curves.

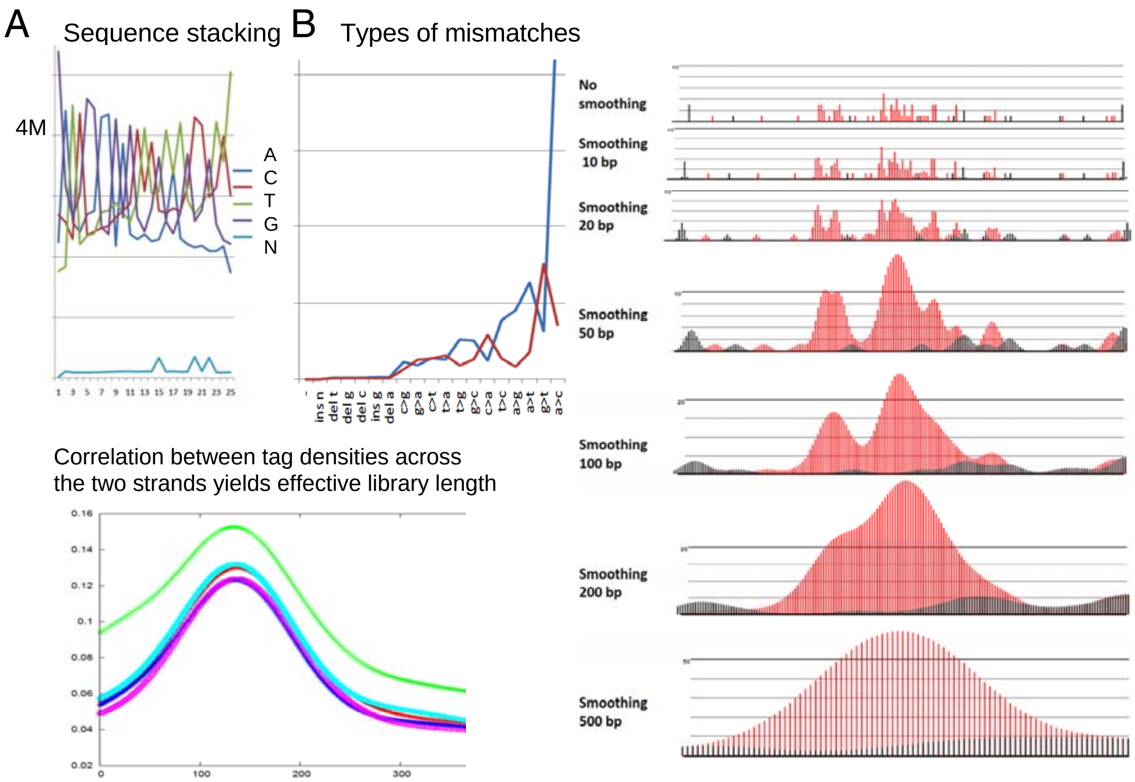
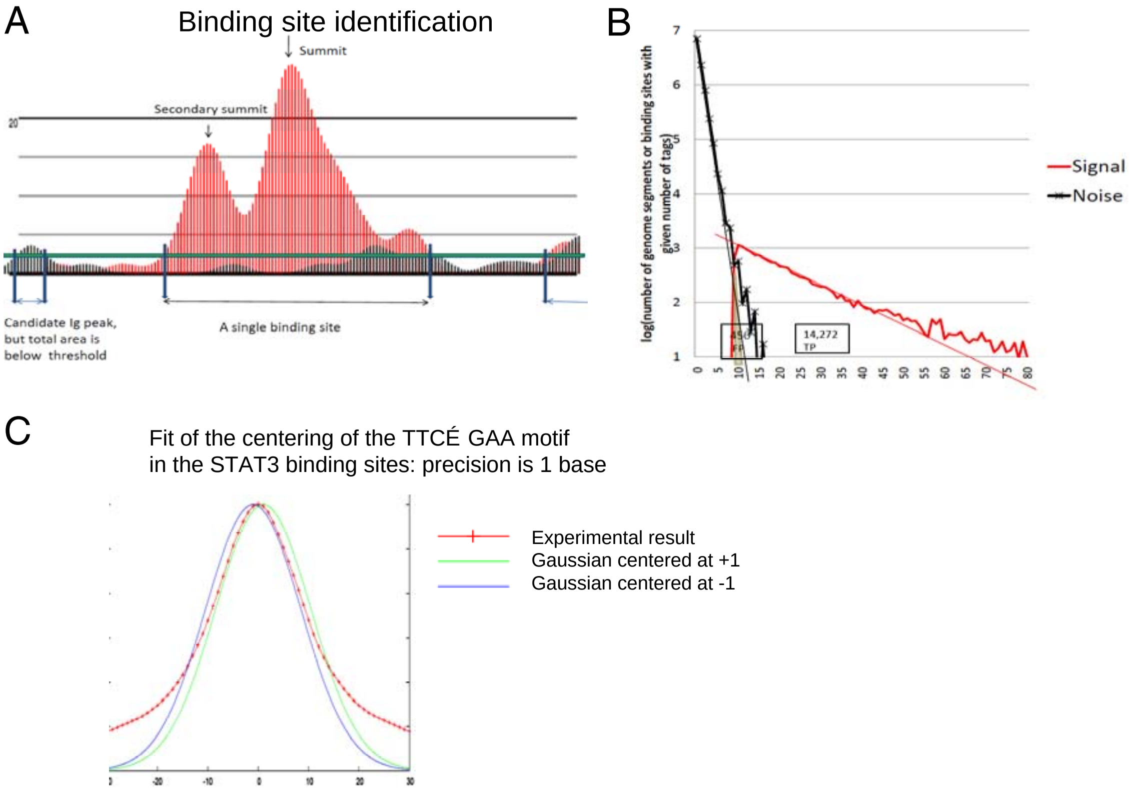

A
Sequence stacking
4M
A
C
T
G
N
1 3 5 7 9 11 13 15 17 19 21 23 25
B
Types of mismatches
-
ins n
del t
del g
del c
ins g
del a
c>g
g>a
c>t
t>a
t>g
g>c
c>a
t>c
a>g
a>t
g>t
a>c
Correlation between tag densities across the two strands yields effective library length
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
100
200
300
No smoothing
Smoothing 10 bp
Smoothing 20 bp
Smoothing 50 bp
Smoothing 100 bp
Smoothing 200 bp
Smoothing 500 bp

Figure S5. Binding site definition and analysis. (A) Identifying binding sites Considering the profile of tags density after Gaussian smoothing of the experimental (red) and control (black) tags, calling a site requires first that the curve emerges above the base threshold (in green), and that the area of the region delimited by the intersection with the base threshold exceeds a minimal area threshold. The summit of the peak is the region where binding is expected to be maximal. Most peaks are simple and appear as a monomodal Gaussians; then the center coincides with the summit. However, the strongest binding sites often are wide and complex, as shown in the illustrated example where one can infer that the protein binds to more than one location within the binding site. In such cases, we usually find binding motifs within short range of both primary and secondary summits. B) Analysis of the signal and noise distributions. In this typical example, 14,272 regions of the genome where IRF4 binds before IL-21 treatment, extending on average over 401 bases, were observed. The red curve (signal) shows the histogram of the number of tags per binding site: the $\log_{10}$ of the number N of sites (y axis) is plotted against the number n of tags (x axis). The threshold to call a binding site is set just above the intersection of the signal and noise trend lines (thin red and black lines), at 9 Solexa tags in this experiment. The average width of the binding sites increases with n, but is 250 bp near this threshold. To calculate the noise curve (steep black line), the genome sequence is deprived of the 14,272 areas containing a binding site, then split into segments of length 250 bp, and the number of tags in each segment is counted. The plot shows the $\log_{10}$ of the number of segments N containing a given number n of tags. A few ‘noise’ segments have a relatively high number of tags, as binding sites are defined after substracting the control immunoglobulin tags from the signal tags. In these semi-log coordinates, the two curves are well approximated by straight lines, with equation $N(n) = N(0)\ a^{-n}$. This log-linear distribution differs markedly from a Poisson

distribution for small and large values of n. The triangle (grey) delimited by the trend lines to the signal and noise curves gives an estimate of the number of false positive binding sites (450), 3% in this experiment. The thresholds chosen in this way, the number of binding sites and estimated false positives are given for all experiments in Supplementary Table 1. (C) Sequence analysis of the binding sites. Genome-wide centering of the GAS motifs in the binding sites is observed. The The experimental curve in red gives the histogram of the distance from the central base of the TTCnnnGAA motif to the maximum of the binding peaks. The central part is well-approximated by a Gaussian distribution with sigma=10 and actually fits between 2 such curves centered at -1 bp and +1bp. Hence, the centering of the motif on the maxima is remarkably precise.


Binding site identification

Fit of the centering of the TTCÉ GAA motif in the STAT3 binding sites: precision is 1 base

Figure S6. Relationship between IL-21-regulated genes and binding of STAT3 and/or IRF4 in functional candidate sites. Chromosomal map of IL-21 regulated genes "within reach" of a binding site. Genes have on average two candidate sites. Shown are genes with at least one STAT3-IRF4 composite site, independent STAT3 and IRF4 sites, only STAT3 site(s) or only IRF4 site(s). IL-21 regulated genes are frequently associated to composite sites that bind both STAT3 and IRF4. Note the paucity of IL-21 sensitive genes on the X chromosome and in a few large chromosomal areas, and their high density clustering in other domains.

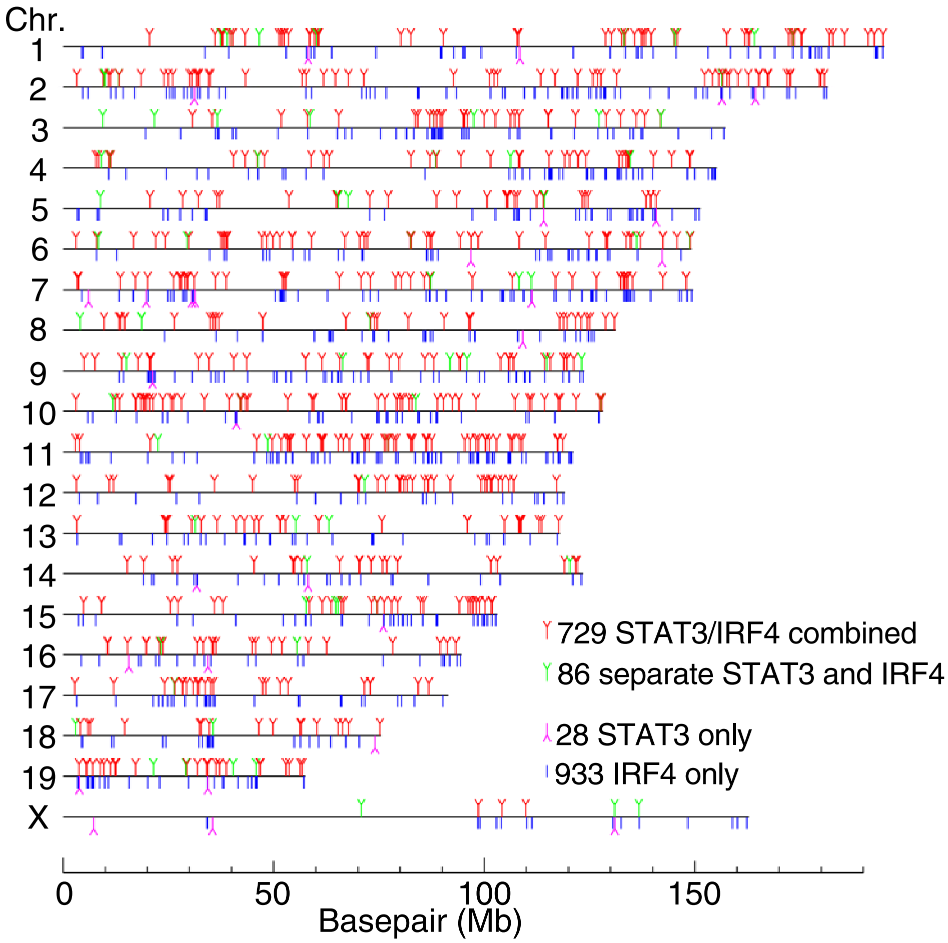
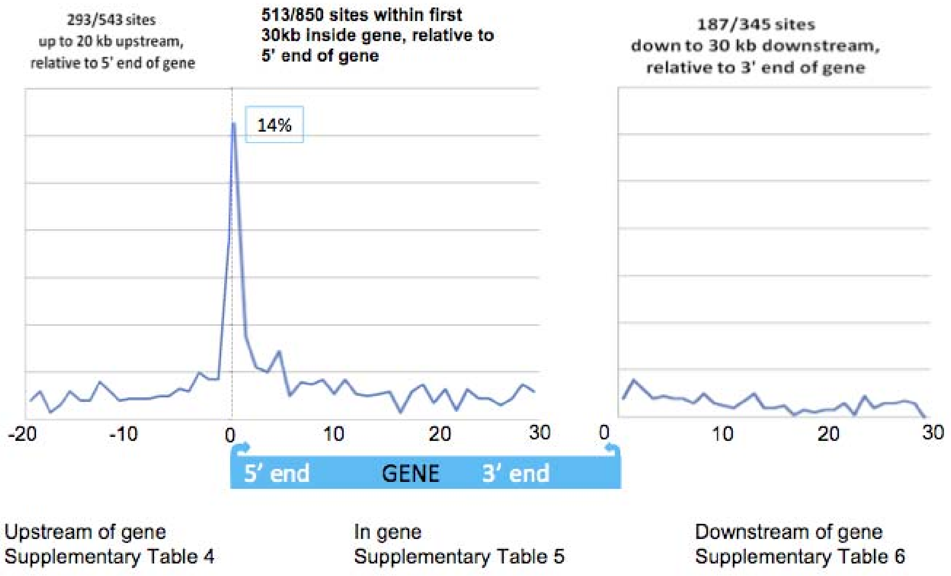
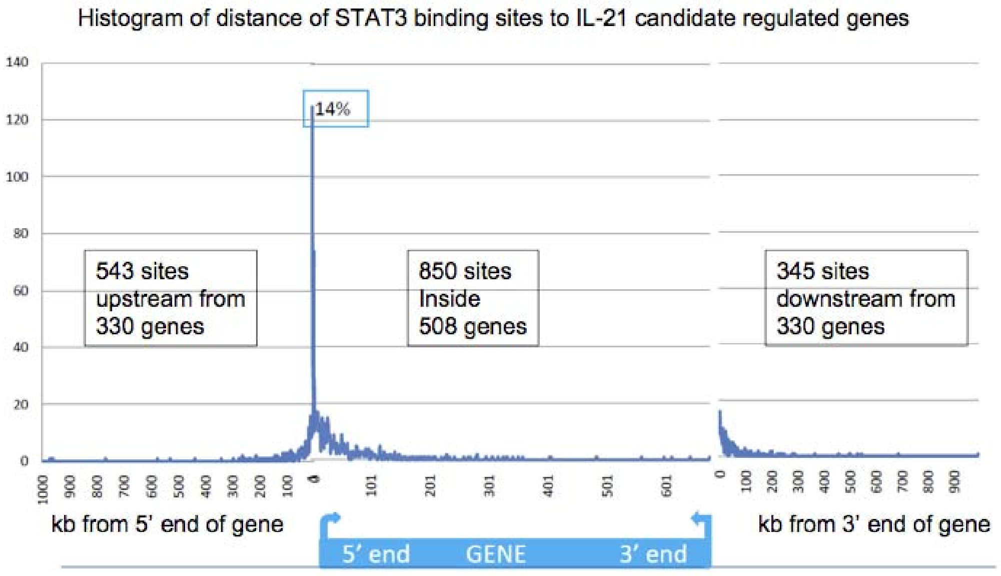


Figure S7. Histogram of distances of STAT3 binding sites to IL-21 candidate regulated genes. The lists of STAT3 sites and their candidate IL-21-regulated genes, mapping in the three sectors depicted here, upstream of genes, inside genes or downstream of genes are given in Supplemental Tables S5, S6, and S7, respectively. (A) and (B) show the histograms at two different scales. The entire region is close to 3 Mb in the top diagram (A) whereas it is limited to 100 kb in the bottom diagram (B). As indicated, there are 543 STAT3 binding sites located upstream of 330 genes, 850 sites located downstream of the transcription start site within 508 genes, and 345 sites that are located downstream of 330 genes, which are best seen in (A). In (B), there is a zoomed in higher resolution at the region of the transcription start site. A total of 14% of the STAT3 binding sites are within 1 kb of the TSS.





Histogram of distance of STAT3 binding sites to IL-21 candidate regulated genes

Figure S8. IRF4 is essential for Tfh differentiation. The proportion of CXCR5$^+$ICOS$^+$CD4$^+$ Tfh cells (A and B) and PNA$^+$B220$^+$ germinal center B cells (C and D) were decreased in *Irf4*$^{-/-}$ mice. Means ± SEM of 2 independent experiments, total combined samples ≥ 4. *, $p < 0.05$ (compared with immunized WT control mice).

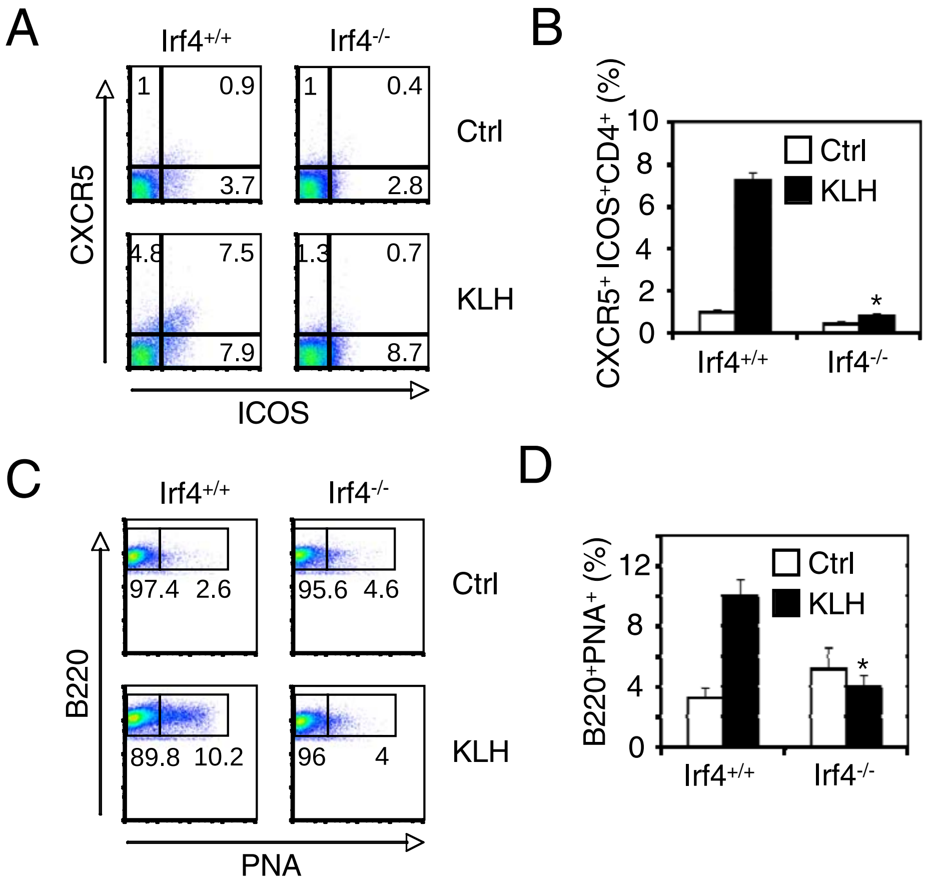


Figure S9. STAT3 and IRF4 binding sequences in the *Socs3*, *Bcl3*, and *Tha1* genes in ChIP-Seq. Genomic sequences of closely positioned STAT3 and IRF4 binding sites in *Socs*3, *Bcl*3, and *Tha*1 genes as shown in Figure 6B. Bold sequences: STAT3 binding motif, underline: IRF4 binding motif.

Socs3

#1 (400 bp)

1178268447:agtgtagagtcagagttagagccgcctcggaggccgcgcgcgcgggtatttacccggccagtacgc
cccgccccccga**ttcctggaa**ctgcccggccggtc**ttcttgtaa**tgtttagtcactactctgcactgaaaggctgtgcgcg
gagggcgagggaggggccgcggagggcgggcttggagctggggcctccaggacccgccgagactcaccgagagg
gagacaaagcgcggcgcgaggctgcccgaccggcgggcgcggcgccagccttggccgagcgttcctggcagcggc
ccctcccccgcgcgctccgcccccaacttctcattcacactttcccccccctccc**ttctaagaa**ggctggtttctggcagag
gcgggggcgtcgcgatgggagc:1178268849

Bcl3

#1 (200 bp)

20401722:tcccggcgcgggcggggttaagcgagggca
gagggagcgaatgat**ttcagagaa**accgtccaggcttcag
ttcccataaacgcgcgggccgcagggggtggggcggggc
tggaggaagtggggacgggaggggcggcccagccctcg
gac**ttccagtaa**caggtctgtggggcggggctgggcttcct
gctcccgggga:20401921

#2 (350 bp)

20402659:tattataagctggagccacacaatgctgg
ctctttctgagtcttgctgccctccacgcagtcaaggctcg
accagaatctctgggatctagacttgaggtcccctgccc
agcctcgctgccttcccctcccctcctccagcctgtgtgaa
gtggggctcgtggctcaggcaagctggggccaggccc
gctgccca**ttccggtaa**gccccagcggaaggggttaa
ggttggagggagcaccgggaggggcaggctgtgagg
agtggaggatgctgacgtgggggaggctgagcagctg
ggctgcggtcaagttgtgcgggaagggatttccccgag
gcggcagctgaggca:20403008

Tha1

#1 (200 bp)

117733272:gtaaccacgtgccacagccagaggcccct
ctgtcagtcctcctgagccacagtccccgccccaggccacc
gaaactcccttgtggtttgaagcccctttcccacccagacctc
acacctgtttttctccct**ttcccagaa**gccccggcaggtgag
gccttgtggtttccctgtggacttcctctgcacatcaccacgga
a:117733471

#2 (200 bp)

117733689:accctcctcagactcctgtcgaggccagg
acctcccagccttgatgtggtgacctgtttacaccaggtgcc
tgctgagagcagtgtagctcagcatctttgaag**ttcctggt**
aatatcttgttgtcaagagattctagatcagccccaatcacc
cctgggtccacaaaggtcaagtgtgatctctggtgtatctct
ggat:117733888

Supplemental References

Holtschke, T., Lohler, J., Kanno, Y., Fehr, T., Giese, N., Rosenbauer, F., Lou, J., Knobeloch, K. P., Gabriele, L., Waring, J. F., *et al.* (1996). Immunodeficiency and chronic myelogenous leukemia-like syndrome in mice with a targeted mutation of the ICSBP gene. Cell *87*, 307-317.

John, S., Robbins, C. M., and Leonard, W. J. (1996). An IL-2 response element in the human IL-2 receptor alpha chain promoter is a composite element that binds Stat5, Elf-1, HMG-I(Y) and a GATA family protein. Embo J *15*, 5627-5635.

Lee, C. H., Melchers, M., Wang, H., Torrey, T. A., Slota, R., Qi, C. F., Kim, J. Y., Lugar, P., Kong, H. J., Farrington, L., *et al.* (2006). Regulation of the germinal center gene program by interferon (IFN) regulatory factor 8/IFN consensus sequence-binding protein. J Exp Med *203*, 63-72.

Lee, C. K., Raz, R., Gimeno, R., Gertner, R., Wistinghausen, B., Takeshita, K., DePinho, R. A., and Levy, D. E. (2002). STAT3 is a negative regulator of granulopoiesis but is not required for G-CSF-dependent differentiation. Immunity *17*, 63-72.

Mittrucker, H. W., Matsuyama, T., Grossman, A., Kundig, T. M., Potter, J., Shahinian, A., Wakeham, A., Patterson, B., Ohashi, P. S., and Mak, T. W. (1997). Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function. Science *275*, 540-543.

Moreno, C. S., Beresford, G. W., Louis-Plence, P., Morris, A. C., and Boss, J. M. (1999). CREB regulates MHC class II expression in a CIITA-dependent manner. Immunity *10*, 143-151.

Polli, M., Dakic, A., Light, A., Wu, L., Tarlinton, D. M., and Nutt, S. L. (2005). The development of functional B lymphocytes in conditional PU.1 knock-out mice. Blood *106*, 2083-2090.

Turner, C. A., Jr., Mack, D. H., and Davis, M. M. (1994). Blimp-1, a novel zinc finger-containing protein that can drive the maturation of B lymphocytes into immunoglobulin-secreting cells. Cell *77*, 297-306.