rVISTA for comparative sequence-based discovery of functional transcription factor binding sites

Gabriela G Loots[1,*], Ivan Ovcharenko[1], Lior Pachter[2], Inna Dubchak[1,3,*] and Edward M Rubin[1]

[1]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [2]Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720, USA. [3]National Energy Research Supercomputing Center, Lawrence Berkeley National Laboratory, CA 94720, USA. *Corresponding authors.

EMAIL: ggloots@lbl.gov; ildubchak@lbl.gov.

ABSTRACT

Identifying transcriptional regulatory elements represents a significant challenge in annotating the genomes of higher vertebrates. We have developed a computational tool, *rVISTA***,** for high-throughput discovery of *cis*-regulatory elements that combines transcription factor binding site prediction and the analysis of inter-species sequence conservation. Here, we illustrate the ability of *rVISTA* to identify true transcription factor binding sites through the analysis of AP-1 and NFAT binding sites in the 1 Mb well-annotated cytokine gene cluster[1] (Hs5q31; Mm11). The exploitation of orthologous human-mouse data set resulted in the elimination of 95% of the 38,000 binding sites predicted upon analysis of the human sequence alone, while it identified 87% of the experimentally verified binding sites in this region.

INTRODUCTION

A major challenge of the post genome-sequencing era is decoding the regulatory networks underlining gene expression. Intensive efforts over several decades have identified numerous regulatory proteins - transcription factors (TF) - whose sequence-specific DNA binding activity is central to transcriptional regulation. The DNA binding specificity of many TFs has been experimentally determined. Most TF bind to short (6-12 basepairs), degenerate sequence motifs that occur very frequently in the human genome. The binding specificities of these factors can be summarized as position weight matrixes (Heinemeyer et al. 1998) that are compiled in various databases such as the *TRANSFAC* database (http://www.biobase.de). Pattern-recognition programs such as *MATCH* or *MatInspector* (Quandt et al. 1995) use these libraries of transcription factor position weight matrixes to identify significant matches in DNA sequences. A major confounding factor in the use of position weight matrixes is that only a small fraction of predicted binding sites are functionally significant, and thus position weight matrixes have proven to be a poor resource for sequence-based discovery of regulatory elements (Pennacchio and Rubin 2001). Proposed strategies for reducing the number of false-positive predictions include clustering of composite sites (Wagner 1997) and using multi-species comparative sequence analysis or phylogenetic footprinting (Gumucio et al. 1996).

The increasing availability of human and mouse orthologous sequences is facilitating large-scale comparative analyses that can be used to improve the computational identification of real transcription factor binding sites (TFBS) (Pennacchio and Rubin 2001). Comparative sequence analysis of orthologous regions from multiple closely related species has proven to be an effective means of identifying evolutionarily conserved sequences and functional noncoding elements (Hardison et al. 1997; Oeltjen et al. 1997; Hardison et al. 2000; Loots et al. 2000). The largest human-mouse comparative analyses have been recently performed for the non-repetitive segments of human chromosome 19 (Dehal et al. 2001) and human chromosome 21 (Frazer et al. 2001) and have provided the initial insight into the global pattern of conservation for the alignment of complete mammalian genomes. The most striking finding from these large-scale comparisons has been the identification of a large number of evolutionarily conserved sequences that bear no resemblance to coding DNA and are embedded in genomic regions void of known genes (Frazer et al. 2001) in addition to being scattered among coding exons (Loots et al. 2000; Dehal et al. 2001). Whereas small-scale sequence comparisons have been extremely useful in identifying functional regulatory elements in the vicinity of known genes, genome-wide comparisons result in an overwhelming number of evolutionarily conserved sequences, and experimentally defining the biological role of these elements represents a major bottleneck.

Currently, we lack adequate computational tools to assist us in characterizing evolutionarily conserved noncoding sequences on a genome-wide scale. Besides coding for proteins, embedded into our genome are DNA sequences involved in vital regulatory functions such as replication, recombination, chromosomal condensation and segregation as well as control elements that determine the expression pattern of genes (Frazer et al. 2001). An important leverage we have in identifying DNA sequences involved in gene regulation is the knowledge that gene expression is regulated by transcription factors that bind selectively to specific DNA sequences, as well as a collection of experimentally defined transcription factor binding sites (TFBS) (Wagner 1997). We have developed a computational tool: *Regulatory VISTA* (*rVISTA*: http://www-gsd.lbl.gov/vista/rVISTA.html) that combines cross-species DNA conservation with TFBS data to overcome some of the limitations associated with single species TFBS predictions and enrich for transcriptional regulatory signals in large genomic intervals. Here, we introduce the *rVISTA* program and illustrate its ability to identify true transcription factor binding sites through the analysis of AP-1 and NFAT binding sites in the 1 Mb well-annotated cytokine gene cluster (Hs5q31; Mm11) (Frazer et al 1997; Loots et al. 2000).

RESULTS

Algorithm

Transcription factor binding site (TFBS) analysis performed by the *rVISTA* algorithm proceeds in four major steps: (1) identification of TFBS matches in the individual sequences, (2) identification of aligned noncoding TFBS, (3) calculation of local conservation across each aligned TFBS, and (4) graphical visualization of noncoding binding sites (Figure 1). Given a global alignment (generated by *AVID*; http://bio.math.berkeley.edu/avid/) and the corresponding sequence annotations as an input, the program calls all available position weight matrices in the *TRANSFAC* database, and independently locates all TFBS matches in the orthologous sequences using the *MATCH* program. The detailed global alignment and the annotation of the two sequences are next used to identify orthologous TFBS that are aligned (≤6 bp shift of the core, ≤1 gap in the TFBS alignment) in noncoding genomic intervals. Consequently, the *Hula-Hoop* component of the *rVISTA* algorithm calculates the DNA conservation for each aligned TFBS as percent identity (%ID) over a 21 bp dynamically shifting window, that centers on a nucleotide inside the TFBS with the maximum %ID. Using the same principle, *rVISTA* calculates the maximum DNA conservation over larger DNA segments (up to 200 bp in 20 bp increments) facilitating the identification of sites present in larger, highly conserved elements. The *rVISTA* algorithm generates two types of outputs: (1) a static data table with detailed statistics for all conserved TFBS and (2) a dynamic web-interactive module that allows the user to customize the data for unfiltered, aligned or conserved TFBS site and graphically visualize them as colored tick marks.

Combinatorial analysis of TFBS using multi-clustering

Detailed molecular analyses addressing the architecture of complex regulatory regions in higher eukaryotes have established that the majority of transcriptional control elements such as enhancers and repressors represent a conglomerate of multiple TFBS that act in concordance to directly modulate the expression patterns of the linked genes. In addition, it has been observed that regulatory elements involved in similar physiological functions, such as the enhancement of liver specific genes (Krivan and Wasserman 2001), are associated with distinct patterns of multiple TFBS. These regulatory regions are frequently present in clusters (Fickett and Wasserman 2000) two or more repeated TFBS, or clusters of two or more adjacent TFBS belonging to different family of regulatory proteins. To be able to perform combinatorial analysis of conserved TFBS, *rVISTA* calculates the distance between neighboring sites and selectively clusters 2 or more sites of the same TF present in regions of user-defined lengths (Fig 2A). Furthermore, *rVISTA* facilitates the identification of evolutionarily conserved elements that harbor multiple clusters of various unrelated TFBS by allowing the user to perform customized combinatorial analysis of TFBS (Fig 2B).

Collection of experimental data and validation of *rVISTA*

To evaluate the biological significance of TFBS data generated by the *rVISTA* algorithm we analyzed ~1 Mb of a well annotated interleukin gene cluster (Hs5q31; Mm11) (*IL-3; IL-4; IL-5; IL-13; IRF-1; GM-CSF*) (Frazer et al. 1997; Loots et al. 2000), plus the intensively characterized interleukin 2 (*IL-2*) promoter region (Hs4q26; Mm3) (Rooney et al. 1995). Cytokines are of particular biomedical importance since they augment the growth and differentiation of T helper cell subsets, and mediate both the protective and pathologic effects of activated T cells. Several of them have been directly implicated in asthma phenotypes and inflammatory disorders (Lacy et al. 2000). As such, much interest has focused on the regulatory mechanisms by which naive helper CD4+ T cells establish their cytokine repertoires, events that are predominantly regulated at the transcriptional level. In addition, cytokine expression is associated with chromatin remodeling such that expression patterns become clonally inherited epigenetic traits (O'Garra and Arai 2000), thereby stabilizing the immune repertoire, and it has been shown that several cytokines, particularly IL-2 and IL-4, can be expressed both mono- and biallelically, consistent with independent regulation of these gene loci (Hollander et al. 1998; Bix and Locksley 1999).

Due to the vast interest in understanding the regulation of these cytokines, we were able to compile a representative collection of 19 experimentally defined AP-1 binding sites and 21 NFAT binding sites (Table 1) from the published data which we used to examine *rVISTA's* ability to identify true binding sites. By analyzing the noncoding DNA of the ~1 Mb human sequence independent of the mouse sequence, the *MATCH* program predicted 23,457 AP-1 and 14,900 NFAT sites using the position weight matrices available in the *TRANSFAC* database (parameters: 0.75/0.8) for these transcription factors. A comparable number of sites were independently predicted for the ~1 Mb mouse sequence. Among the predicted AP-1 and NFAT sites for the human sequence were also included 17 of the 19 functional AP-1 sites and 19 of the 21 functional NFAT sites. The omitted AP-1 and NFAT functional sites failed to meet the *TRANSFAC* default parameters. Subjecting the orthologous human and mouse sequences to *rVISTA* analysis reduced the total number of predicted AP-1 and NFAT sites in the ~1 Mb human sequence by greater than 95%, identifying 1114 conserved AP-1 and 734 conserved NFAT sites. In sharp contrast, *rVISTA* identified 16 of the 19 AP-1 and 19 of the 21 functionally characterized NFAT sites, establishing a strong correlation between sequence conservation and functional TFBS.

Cytokine promoter analysis

Although the loci encoding these cytokines are not homologous at the nucleotide or amino acid level, they are considered a gene family because of their overlap in biological activities, and secondary and tertiary structural similarities. Recent advances have also revealed similarities in their expression pattern, and activation by the same T-cell specific transcription factors. In addition to AP-1 and NFAT, the GATA-3 TF has also been implicated in the transcriptional control of these T_h2-specific cytokines[10]. GATA-3's direct involvement in gene activation has been extensively demonstrated for the *IL-4*and *IL-5*promoters[11, 12] and has been postulated for the activation (or repression) of all the interleukin genes present in this interval[11]. Based on GATA-3′s predicted binding to upstream regions of cytokine genes, we hypothesized that there should be an increased distribution of GATA-3 site across the 6 cytokine promoters compared with the promoters of the 16 non-T_h1/ T_h2 expressing genes in this region. To test this hypothesis, we determined the GATA-3 site distribution for the 2 kb promoter region of all 22 annotated genes in this interval. Due to the highly degenerate nature of the GATA binding profile that is recognized by all members of the GATA-family[13], *TRANFAC* predicted an average of 50 GATA-3 sites per promoter that were evenly distributed across both cytokine and non-cytokine gene promoters. In contrast, the *rVISTA* analysis dramatically reduced the total number of GATA-3 sites per promoter, and most importantly resulted in an increased representation of GATA-3 sites in cytokine promoters (Figure ***3a***). On average *rVISTA* detected 8 conserved GATA-3 sites per cytokine promoter while yielding only 2 conserved GATA-3 sites per non-cytokine promoter. In addition, the functionally known GATA-3 sites in both the *IL-4* and *IL-5* promoters were among the highly conserved sites identified by *rVISTA*. Since functional GATA-3 sites are present in pairs (Supplemental Table), we next analyzed the distribution of GATA-3 sites clustered (2 or more sites present within 60 bp regions) (Figure ***2b***). By clustering the conserved GATA-3 sites, we observed, a further enrichment of GATA-3 sites in the cytokine promoters. In each cytokine promoter there were an average of 6 GATA-3 clustered sites, while no GATA-3 clustered sites were noted in non-cytokine promoters, supporting the hypothesis that GATA-3 plays an important role in the regulation of all the interleukin genes present on human 5q31.

DISCUSSION

Annotating the noncoding portion of the human genome remains among the greatest challenges of the post-sequencing era. Even though cross-species comparisons can be used to identify functional elements, in depth characterization of these elements represents a major bottleneck in identifying the biological roles of these DNA sequences while analyzing whole genomes. To assist genomically-informed studies of noncoding sequences while analyzing large genomic intervals and to enhance the detection process of putative regulatory signals involved in transcriptional control of eukaryotic genes the algorithm we have developed identifies evolutionarily conserved TFBS, and detects clusters of binding sites for known transcription factors. By performing an unbiased analysis of the distribution of NFAT and AP-1 binding sites across ~1Mb of human/mouse orthologous region, we were able to show that while *rVISTA* reduces more than 95% of the predicted TFBS, it recognizes 87% of the functionally characterized AP-1 and NFAT sites. The compiled position weight matrices available in the *TRANSFAC* database pose a major limitation in the *rVISTA* analysis, since the computational approach described relies on the available DNA binding profiles of known transcription factors. 3 of the 4 experimentally defined sites that failed to meet the default sequence match criteria in the initial analysis of the human sequence were conserved greater than 80%, while the remaining site was conserved at 76% (Table 1). Of the 36 experimentally defined sites recognized by the position weight matrices available in the *TRANSFAC* database, only one AP-1 site was lost during the *rVISTA* analysis, but it is worth mentioning that this site was also conserved at ~71% identity. To overcome some the limitation set by the available position weight matrices, we are currently implementing an additional option to the rVISTA algorithm which will allow users to build their own position weight matrices and search for consensus sequences independent of the *TRANSFAC* database.

While identification of conserved binding sites can also be easily achieved by phylogenetic footprinting, the greatest strength of the *rVISTA* algorithm are the clustering modules and the user-defined customization of visualized sites. Through the use of a global alignment *rVISTA* takes into account the linear structure of DNA. Properties related to protein-protein interaction, chromatin structure, as well as clusters of multiple unique sites that have been reshuffled in one of the human or the mouse genome and have lost their positional linearity are not addressed. Also, clustering does not take into account the spacing between sites, but rather counts the number of adjacent sites of a given TF spanning DNA segments of specified length. In the future we will consider the interaction between TF binding to adjacent sites and incorporate this variable into additional analytical modules of the *rVISTA* algorithm.

Our analysis of the AP-1 and NFAT TFBS in the cytokine gene cluster illustrates the effectiveness of the *rVISTA* algorithm in eliminating many false positives while retaining the majority of experimentally verified sites. By exploiting the orthologous sequence data from this region we were able to prioritize, based exclusively on sequence analysis, a limited number of GATA-3 sites that were found exclusively in the proximity of genes predicted to be GATA-3 responsive. With the increasing availability of sequence data for genomes of multiple organisms, *rVISTA*'s ability to utilize comparative data makes it particularly suited to perform large-scale genome-wide analyses to identify transcription factor binding sites with a high likelihood of being true.

METHODS

rVISTA is implemented as a publicly open and free web-based tool [http://www-gsd.lbl.gov/vista/rVISTA.html] that requires an alignment file and optional sequence gene annotation files for user input. The *rVISTA* analysis tool consists of four major modules: (1) motif recognition, (2) identification of aligned TFBS (3) conservation analysis and (4) visualization of TFBS. The system units are implemented using the C++ computer language equipped with Web user-interactive interface written in *Perl*. *rVISTA* utilizes the alignment produced by the *AVID* program [http://bio.math.berkeley.edu/avid/] for the conservation analysis. The following methodological scheme was implemented as a core of the *rVISTA* tool. Initially, for the chosen set of transcription factor matrices with given thresholds the extraction of TFBS coordinates in both homologous sequences is performed. The locally installed *TRANFAC* 5.2 library and the *MATCH* program from Biobase, Inc. [http://www.biobase.de] are used at this step. Next, by using the pairwise alignment for the two orthologous sequences *rVISTA* identifies identical TFBS that are aligned. An aligned TFBS is allowed to have a maximum of six nucleotides shift in its aligned core, and one gap present across the entire local alignment. The conservation analysis module contains one major unit, the *Hula-Hoop*, which is designed to analyze the local DNA conservation of aligned TFBS to eliminate aligned sites present in regions of weak conservation. A fixed-size window is being shifted through all the position of an aligned TFBS, while the entire sequence spanning the TFBS is permanently enclosed by the shifting window. The percent identity is calculated at every point in the DNA window, and the maximum percent identity is assigned for the binding site (DNA window of 21 nucleotides). This unit accounts for the local conservation, allowing the identification of local highly conserved TFBS.

The visualization module is a web-based tool that post-processes the *rVISTA* output. One unit of the program eliminates redundancy. Overlapping TFBS matches (within 3 bp from each other) belonging to the same family of regulatory proteins are considered to be an identical match. A second unit of the program measures the distance between adjacent matches belonging to the same TF family, and allows the user to selectively cluster TFBS into groups of *x* number of sites over *y* base pair length. The clustering parameters are user-defined and assigned independently for every family of TF. Any combination of unfiltered, aligned or conserved TFBS with customized clustering for the selected set of TFs are interactively visualized as a ‘tick-plot’ track overlaid on the conservation *VISTA*-type track and the gene annotation track.

Statistical Modulation of GATA sites across Promoters

In order to analyze the specificity of the method developed we performed a statistical modulation for the distribution of predicted GATA sites and compared it with the distribution of the observed GATA sites in the promoter regions of (2kb upstream of the 5'UTR). GATA site redundancy was excluded prior to analysis. Site conservation criteria was chosen for >80% identity over a window of 21 bp, and GATA site clustering was performed for 2+ neighboring GATA sites over a <60 bp region. The expected number of conserved GATA sites in a promoter under consideration, *i* was calculated as follows:

$$\mathrm{nGATA}_i = l_i/d.$$

where d=41 bp and is the average distance between two non-redundant GATA sites in the human sequence is. The probability of a given nucleotide to be a starting position for a GATA site is $1/d$, and l_i is the number of nucleotides inside a promoter region conserved >80%. The values obtained represent the upper boundary for the expected number of aligned and conserved GATA sites in a given promoter. To obtain a more accurate expected value for the predicted GATA sites, we also considered the fact that a conserved sites is not always an aligned site. Any given GATA site could either have no corresponding binding site in the second sequence that is aligned to it (because the identity is not usually 100%) or the TFBS alignment could exceed the number of total gaps allowed. We approached this problem by introducing the scaling parameter σ, which is a probability for the clustered site to be aligned at the same time. σ is less or equal to 1 and is approximated to be constant for all the promoter regions. Then,

$$\mathrm{nGATA}_i = \sigma * l_i/d.$$

The estimation for the value of σ was calculated based on the normalization approach that the total number of clustered GATA sites expected and observed should be the same. It is worthwhile to mention that the normalization was done based only on the promoters of non-cytokine genes. It let us to compare the distribution of the GATA conserved sites in cytokine gene promoters versus the statistical estimation of the GATA conserved sites in non-cytokine gene promoters [Figure 3a].

Similarly we estimated the number expected conserved and clustered GATA sites. This time, the length of the conserved and clustered segment of the promoter was obtained by checking all the possible paired coordinates in the conserved regions of every promoter and obtaining the fraction of sites which are closer than 60 bps but greater than 4 bp apart from each other. The conserved and conserved-and-clustered σ values were found to be equal to 0.23 and 0.19, respectively. The close value of σ in both cases supports the hypothesis about the possibility to introduce a single scaling parameter σ, which is the same for all the promoters.

ACKNOWLEDGEMENTS

We are grateful to Jody Schwartz, Alexander Fabrikant, Michael Brudno and Moshe Malkin for technical assistance. We thank Mark Biggin, Michael Eisen and all the members of the Rubin laboratory for their insightful comments on the manuscript.. This work was supported by the Program for Genomic Applications (PGAs) funded by the National Heart, Lung, and Blood Institute (NHLBI/NIH); GGL was supported by the Department of Energy- Alexander Hollaender Fellowship.

REFERENCES

Bix M, Locksley RM. Independent and epigenetic regulation of the interleukin-4 alleles in CD4+ T cells. (1998) *Science*. 281(5381):1352-4.

Dehal P, Predki P, Olsen AS, Kobayashi A, Folta P, Lucas S, Land M, Terry A, Ecale Zhou CL, Rash S, Zhang Q, Gordon L, Kim J, Elkin C, Pollard MJ, Richardson P, Rokhsar D, Uberbacher E, Hawkins T, Branscomb E, Stubbs L. Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* 2001 Jul 6;293(5527):104-11

Fickett JW and Wasserman WW. Discovery and modeling of transcriptional regulatory regions. (2000) *Curr Opin Biotechnol*. 11(1):19-24.

Frazer KA, Ueda Y, Zhu Y, Gifford VR, Garofalo MR, Mohandas N, Martin CH, Palazzolo MJ, Cheng JF, Rubin EM. Computational and biological analysis of 680 kb of DNA sequence from the human 5q31 cytokine gene cluster region. (1997) *Genome Res*. 7(5):495-512.

Frazer KA, Sheehan JB, Stokowski RP, Chen X, Hosseini R, Cheng JF, Fodor SP, Cox DR, Patil N. Evolutionarily conserved sequences on human chromosome 21. Genome Res. 2001 Oct;11(10):1651-9.

Gumucio DL, Shelton DA, Zhu W, Millinoff D, Gray T, Bock JH, Slightom JL, Goodman M. Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes. (1996) *Mol Phylogenet Evol*. 5(1):18-32.

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA.

Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. (1998) *Nucleic Acids Res*. 26(1):362-7.

Hollander GA, Zuklys S, Morel C, Mizoguchi E, Mobisson K, Simpson S, Terhorst C, Wishart W, Golan DE, Bhan AK, Burakoff SJ. Monoallelic expression of the interleukin-2 locus. Science. 1998 Mar 27;279(5359):2118-21.

Krivan W and Wasserman WW. A predictive model for regulatory sequences directing liver-specific transcription. (2001) Genome Res 11(9):1559-66.

Lacy DA, Wang ZE, Symula DJ, McArthur CJ, Rubin EM, Frazer KA, Locksley RM. Faithful expression of the human 5q31 cytokine cluster in transgenic mice. (2000) *J Immunol.* 164(9):4569-74.

Lee HJ, O'Garra A, Arai K and Arai N. Characterization of cis-regulatory elements and nuclear factors conferring Th2-specific expression of the IL-5 gene: A role for a GATA-binding protein. (1998) *J Immunol* Mar 1;160(5):2343-52.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. (2000) *Science*. 288(5463):136-40.

Merika M, Orkin SH. DNA-binding specificity of GATA family transcription factors. (1993) *Mol Cell Biol*. 13(7):3999-4010.

O'Garra A and Arai N. The molecular basis of T helper 1 and T helper 2 cell differentiation. (2000) *Trends Cell Biol* 10(12):542-50.

Pennacchio LA and Rubin EM. Genomic strategies to identify mammalian regulatory sequences. (2001) 2(2):100-9.

Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T. MatInd and MatInspector - New fast and versatile tools for detection of consensus matches in nucleotide sequence data. (1995) *Nucleic Acids Res.*, 23:4878-4884.

Ranganath S, Murphy KM. Structure and specificity of GATA proteins in Th2 development. (2001) *Mol Cell Biol* 21(8):2716-25.

Rooney JW, Sun YL, Glimcher LH, Hoey T. Novel NFAT sites that mediate activation of the interleukin-2 promoter in response to T-cell receptor stimulation. (1995) *Mol Cell Biol*. 15(11):6299-310.

Wagner A. A computational genomics approach to the identification of gene networks. (1997) *Nucleic Acids Res*. 25, 3594-3604.

Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S. The TRANSFAC system on gene expression regulation. (2001) *Nucleic Acids Res*. 29(1):281-3.

Zheng W-P, Flavell RA. The transcription factor GATA-3 is necessary and sufficient for TH2 cytokine gene expression in CD4 T Cells. (1997) *Cell* 89:587-596.

.

Figure 1. Visualization of transcription factor binding sites. *rVISTA* analysis of AP-1 and NFAT sites across 30 kb genomic interval from the *GMCSF-IL-3* region ***(a)***. The total AP-1 and NFAT *TRANSFAC* matches (0.75/0.8) predicted for this region are depicted in blue, the human-mouse conserved ones (>80% over 21 bp) are in green. *rVISTA* analysis of the IL-5 promoter identified two (E1 and E2; Supplemental Table) functionally characterized NFAT-AP1composite sites ***(b)*** and one experimentally defined GATA-3 cluster ***(c)*** indicated by black arrows.

Figure 2. Distribution of conserved GATA-3 binding sites across the 22 promoter regions (2 kb upstream of 5'utr) of all annotated genes from the 1Mb cytokine gene cluster (Hs5q31; Mm11). Cytokine genes are marked by arrows, gray bars indicate observed GATA-3 sites, while open bars represent predicted GATA-3 sites due to random distribution. Random distribution was estimated based on the frequency of GATA-3 sites across the 1Mb human sequence and the DNA conservation of each promoter. ***(a)***, conserved individual GATA-3 sites. ***(b)***, conserved GATA-3 present in clusters (2 or more conserved sites enclosed in a 60 bp DNA fragment).