OPEN ACCESS Freely available online

PLoS one

Prediction and Experimental Validation of Novel STAT3 Target Genes in Human Cancer Cells

Young Min Oh[1☯], Jong Kyoung Kim[2☯], Yongwook Choi[1], Seungjin Choi[2*], Joo-Yeon Yoo[1*]

1 Department of Life Sciences, Pohang University of Science and Technology, Pohang, Republic of Korea, **2** Department of Computer Science, Pohang University of Science and Technology, Pohang, Republic of Korea

Abstract

The comprehensive identification of functional transcription factor binding sites (TFBSs) is an important step in understanding complex transcriptional regulatory networks. This study presents a motif-based comparative approach, STAT-Finder, for identifying functional DNA binding sites of STAT3 transcription factor. STAT-Finder combines STAT-Scanner, which was designed to predict functional STAT TFBSs with improved sensitivity, and a motif-based alignment to minimize false positive prediction rates. Using two reference sets containing promoter sequences of known STAT3 target genes, STAT-Finder identified functional STAT3 TFBSs with enhanced prediction efficiency and sensitivity relative to other conventional TFBS prediction tools. In addition, STAT-Finder identified novel STAT3 target genes among a group of genes that are over-expressed in human cancer cells. The binding of STAT3 to the predicted TFBSs was also experimentally confirmed through chromatin immunoprecipitation. Our proposed method provides a systematic approach to the prediction of functional TFBSs that can be applied to other TFs.

Citation: Oh YM, Kim JK, Choi Y, Choi S, Yoo J-Y (2009) Prediction and Experimental Validation of Novel STAT3 Target Genes in Human Cancer Cells. PLoS ONE 4(9): e6911. doi:10.1371/journal.pone.0006911

Editor: Sridhar Hannenhalli, University of Pennsylvania School of Medicine, United States of America

Received April 2, 2009; **Accepted** August 3, 2009; **Published** September 4, 2009

Funding: This work was supported by grants from the Korea Science and Engineering Foundation (KOSEF) grant funded by the MEST (R01-2008-000-20721-0) and to the National Core Research Center for Systems Bio-Dynamics (R15-2004-033). J. K. Kim is supported by a Microsoft Research Asia fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: jyoo@postech.ac.kr (JY); seungjin@postech.ac.kr (SC)

☯ These authors contributed equally to this work.

Introduction

The ability of any biological system to properly respond to stimuli heavily depends on biochemical cascades of signaling pathways that culminate in the activation of transcription factors (TFs) and the subsequent alteration of gene expression patterns [1]. Information about which genes need to be expressed in a specific cell type at any given time is believed to be encoded in the genome. The molecular machinery used to interpret such genetic information has evolved to ensure the accuracy and specificity of gene regulation. Transcription is a multi-step process requiring the concerted action of many proteins. Transcriptional activators and repressors bind in a sequence-specific manner to promoters or enhancers of target genes. They govern the recruitment of trans-activators, chromatin modifiers, and general transcription factors, including RNA polymerase II, to regulate gene expression [2,3].

Whole genome approaches to measure genome-wide expression patterns have divulged groups of genes that are co-regulated to exert spatially and temporally controlled cellular responses [4]. Identifying the responsible regulatory modules that govern the coordinated actions of combinatorial transcription factors is crucial for understanding the regulatory circuits of biological processes [5]. For this purpose, computational tools have been developed to aid in the identification of transcription factor binding sites (TFBSs) in the promoters of the co-regulated genes [6,7,8]. These computational approaches can be divided into two classes: (1) pattern detection and (2) pattern matching. Pattern detection, also known as de novo motif discovery, finds putative binding sites for unknown TFs that are over-represented in the promoters of co-regulated genes. If the binding specificity of a TF is already known, pattern matching methods are preferred [9]. In the pattern matching approach, DNA sequence information of TFBSs is expressed as a position weight matrix (PWM), which can be used to score potential regulatory sites within a statistical framework [10]. However, because DNA binding sites for TFs are generally short and degenerate, this method is prone to high false positive prediction rates [11].

Based on the observation that conserved non-coding DNA sequences are often important for the regulation of biological functions, cross-species sequence comparisons have been actively integrated to distinguish functional and non-functional TFBSs [12,13,14]. The act of incorporating the evolutionarily conserved sequence information in the regulatory regions filters out the non-conserved TFBSs, thereby greatly reduce the false positive prediction rate [15,16,17,18,19]. Although this approach has been successfully applied to increase the predictive power of motif finding, it is highly sensitive to the algorithm used for sequence alignment and the accuracy of annotated transcriptional start site (TSS) information. Therefore, it has been reported that sequence-based promoter alignments often fail to detect short or degenerate regulatory elements, when evolutionary divergent promoter sequences are aligned [12,17]. To overcome these limitations, an alignment-free algorithm based on network-level conservation has also been suggested [20].

Signal transducer and activator of transcription 3 (STAT3) belongs to the STAT family of transcription factors, which is activated by Interleukin-6 (IL-6) and related cytokines, such as

IL-10, Oncostatin M (OSM), and leukemia inhibitory factor (LIF) [21]. Thus far, seven mammalian STATs (1, 2, 3, 4, 5a, 5b, and 6) have been identified. They all possess a DNA binding domain, an SH2 domain for dimerization, and a C-terminal trans-activation domain [22]. Upon stimulation with extracellular ligand, activated STAT3 forms homodimers or heterodimers with another STAT family member, STAT1, then translocates into the nucleus and binds to cognate regulatory elements in the promoters of STAT-responsive genes. Accumulating evidences suggest that STAT3 also associates with other transcription factors to form enhanceosome complexes in the promoter regions of target genes and controls cooperative gene induction [23,24,25]. STAT3 is involved in diverse cellular responses, including cellular differentiation, survival, stem cell renewal, wound healing and systemic inflammation; this has been proven by the phenotypes of genetically modified STAT3 mutant mice [22,26,27,28,29]. It has been found that STAT3 participates in carcinogenesis, and that the ectopic expression of a constitutively active form of STAT3 (STAT3-C) induces tumor formation in nude mice [30]. Furthermore, the expression of constitutively-active STAT3 has been observed in various types of human cancer including multiple myeloma, colon, ovary, liver, lung, head, and neck cancers [31]. While the regulatory and general trans-activation mechanisms of STAT3 have been thoroughly studied, not too much effort has been made towards the identification of direct target genes of STAT3. The identification of those target genes is crucial for mediating the diverse biological effects of STAT3 signaling.

To characterize STAT3-mediated transcriptional programs, we have developed a computational framework designed to predict STAT3 TFBSs with improved sensitivity and low false positive rate. Through the integration of the microarray data obtained from the STAT3 activation condition and the TFBS prediction tools, we attempted to identify novel STAT3 target genes. Using our STAT-Finder program, we identified eight novel STAT3 target genes among a group of genes that are highly expressed in cancer cells. These were then confirmed through chromatin immunoprecipitation.

Results

Overview of STAT-Finder

To identify direct STAT3 target genes, we developed a computational framework that predicts functional TFBSs of STAT3 with increased sensitivity and low false positive rate. Our framework, STAT-Finder, was constructed based on two computational components, a TFBS scanning program (STAT-Scanner) and a motif-based alignment program (Figure 1). STAT-Scanner was designed to increase the sensitivity for detecting functional STAT3 TFBSs. A currently available STAT3-specific PWM of TRANSFAC database [32], V$STAT3_01, frequently

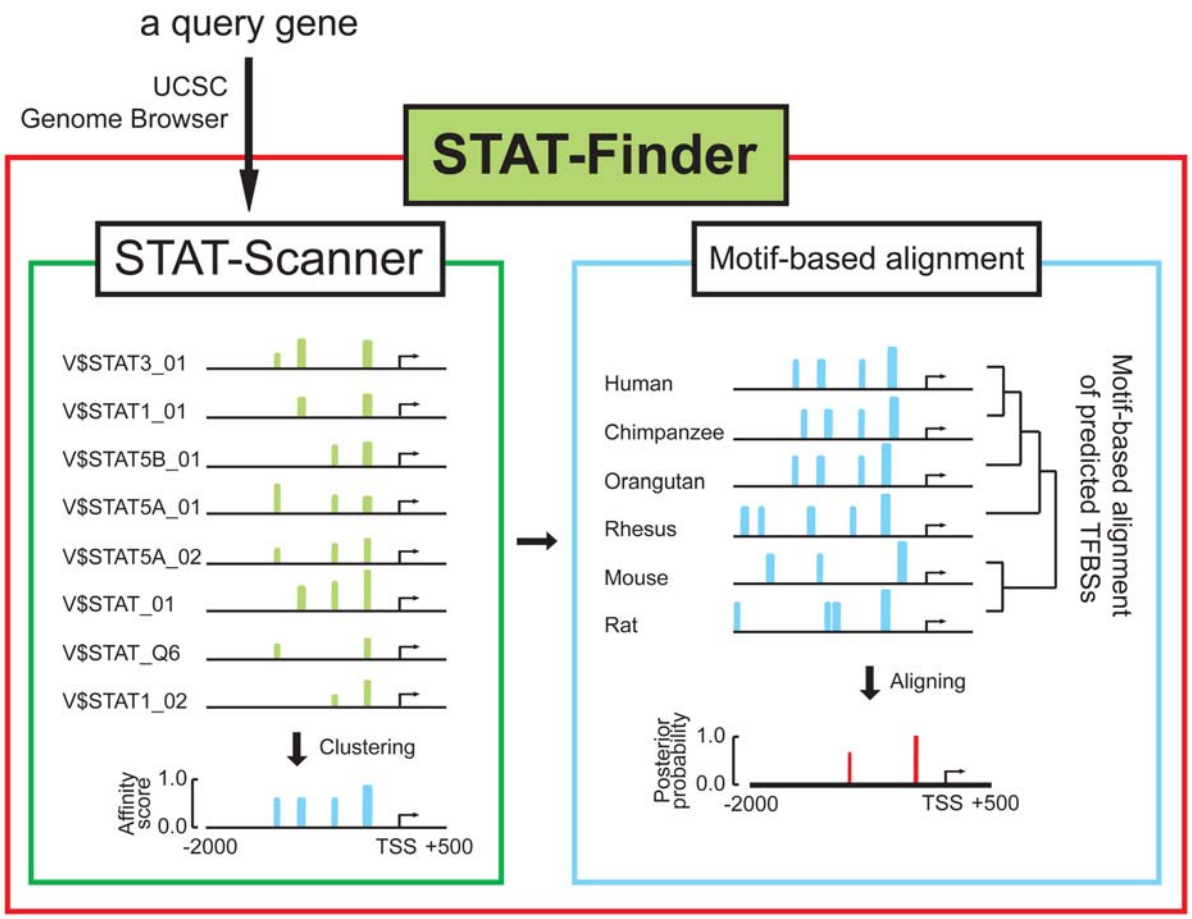

Figure 1. An overview of STAT-Finder. STAT-Finder has two components: The first module, STAT-Scanner, takes a set of six orthologous mammalian promoter sequences as input. Each promoter sequence is searched to mark putative TFBSs using the modified 8 STAT-related PWMs. Binding affinity scores of predicted TFBSs are calculated based on the *P*-values, and a sequence of affinity scores is generated for each promoter. The second module progressively aligns the score sequences and calculates posterior probability to evaluate the degree of motif conservation.
doi:10.1371/journal.pone.0006911.g001

fails to detect experimentally proven STAT3 binding sites (data not shown). For improved predictive power, STAT-Scanner was therefore designed to use combined PWMs of binding specificity similar to STAT3. Although STAT family members have different physiologic functions and regulate distinct sets of target genes, the targets of individual STAT proteins sometimes overlap, and DNA sequences recognized by STAT family members are similar [21,22,23].

For unbiased identification of the PWMs that share sequence similarity with the STAT3-specific PWM, V$STAT3_01, a total of 565 PWMs derived from vertebrate TRANSFAC database [32] were clustered based on their motif similarity (Figure S1). The motif similarity was defined as the *P*-value of the gapped alignment between the two PWMs based on the Kullback-Leibler divergence [33] (See Methods). Total numbers of PWM clusters increased with stringent *P*-value cut-off, reaching maximum cluster numbers of around 10^{-16} *P*-value (Figure S1A). With the *P*-value cut-off of 10^{-7}, PWMs assigned for the STAT family members were found in the same cluster. It is noteworthy that PWM clustering did not reveal any non-STAT PWMs that were similar enough to include nor were there any STAT PWMs that were distinctly different (Figure S1B). We chose among them eight PWMs from the STAT family members with high PWM quality scores (>0.6), where each quality score was calculated using the method proposed by Rahmann et al. [34]. The relevance of the selected PWMs for detecting known STAT3 TFBS has been evaluated in the previously identified STAT3 target genes [35] (Figure S2).

To minimize false positive predictions, results from STAT-Scanner were then analyzed using the comparative motif-based alignment tool (Figure 1). This method finds conserved binding sites within the orthologous promoters of six mammalian species by comparing multiple sequences. Within a probabilistic framework, STAT-Finder then evaluates the posterior probabilities of TFBSs as predicted by STAT-Scanner by assigning higher prior probabilities on conserved sites over non-conserved ones.

Validation of STAT-Scanner

We first compared the performance of STAT-Scanner with the most practical TFBS prediction tools, MATCH 2.7 [36] and MotifLocator [37]. For this purpose, we collected positive genes with experimentally proven STAT3 binding sites in their promoter regions through literature mining and TRED search (http://rulai.cshl.edu/ TRED) [38]. Resulting information on the 22 reference sequences are listed in Table S1. Genomic DNA sequences spanning from 2,000 bp up-stream to 500 bp down-stream of the annotated TSS of each gene were used as input promoter sequences. Prediction of the true positive TFBSs was then plotted as a function of the total predicted TFBS count for different cut-off values. As shown in Figure 2A, STAT-Scanner, which uses combined STAT3-related PWMs, outperforms MATCH and MotifLocator, both of which use the representative STAT3 PWM (V$STAT3_01). We believe the enhanced predictive power of STAT-Scanner was partly due to the usage of combined STAT3-related PWMs, especially since the predictive power of MotifLocator also increased when combined PWMs were used (Figure S3).

We also evaluated the performance of STAT-Scanner using genome-wide STAT3 binding data obtained using embryonic stem cells [39]. Among the 461 genes with STAT3 binding peaks in the 2.5 kb promoter regions, 412 have been accurately predicted by STAT-Scanner to have at least one STAT3 TFBS (Figure 2B). The overall performance of STAT-Scanner was better than those of both MATCH and MotifLocator, as the detection of the same number of true binding sites was achieved by both with significantly lower total numbers of predicted sites. Although MATCH and MotifLocator performed similarly to STAT-Scanner in detecting about 50% of true STAT3 TFBSs, the latter outperforms both by accurately predicting the remaining true sites. We believe this is partly due to the usage of combined STAT-related PWMs which has the capability to enhance the performance of MotifLocator, albeit less than the enhancement for STAT-Scanner, with combined data derived from multiple PWMs (Figure S4). The relative performance of both methods is

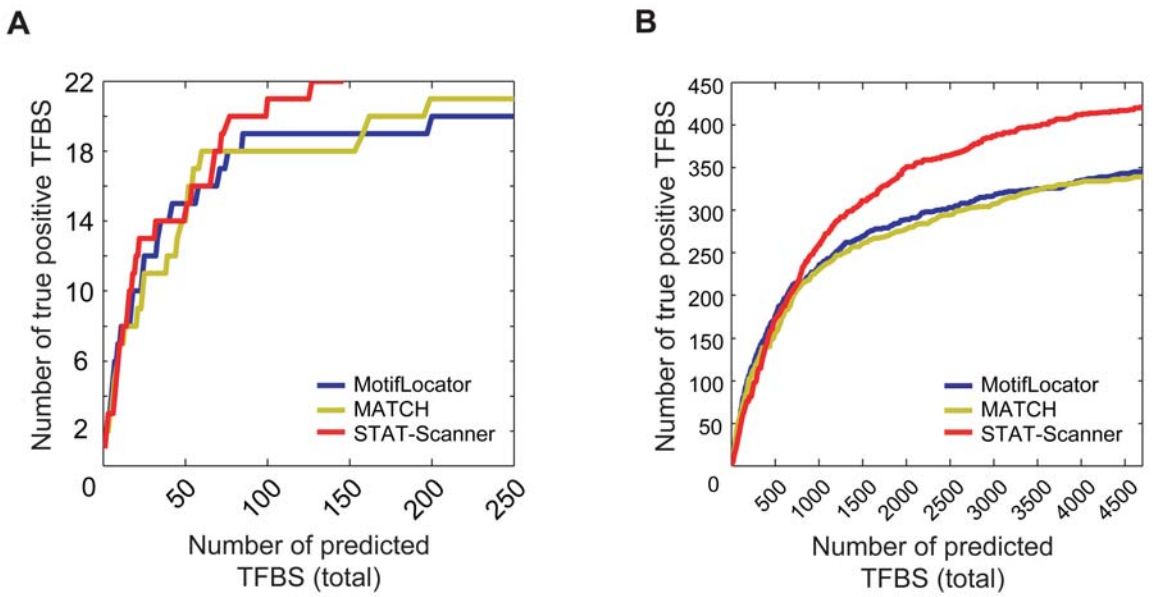

Figure 2. Performance comparison of the STAT3 TFBS prediction tools. Curves for the changes of the number of true positive TFBSs detected using MotifLocator (V$STAT3_01), MATCH (V$STAT3_01), or STAT-Scanner, as a function of total number of predicted TFBSs (A) in the reference set of 22 STAT3 target genes (Table S1) and (B) in the genome-wide STAT3 ChIP-Seq dataset [39].
doi:10.1371/journal.pone.0006911.g002

low compared to that of STAT-Scanner; this can be explained by the fact that their scores on the predicted sites are not directly comparable among different PWMs, thus showing the importance of our scoring scheme in integrating matches to different PWMs. These results also indicate that overlapping PWMs with similar binding specificity are critical to the development of improved strategies to detect functional TFBSs of STAT3 with high predictive accuracy.

Features of the functional STAT3 TFBS

The ultimate goal of computational prediction is to detect functional TFBSs with a high degree of confidence. To filter out the false positive TFBSs with high affinity scores, we examined various functional constraints such as evolutionary conservation and genome structure of predicted STAT3 TFBS regions. Sequence conservation among multiple species has been proven to constrain functional TFBS [16,17,40]. Therefore, we first evaluated the distribution of multispecies conservation scores (PhastCons score) [41] and regulatory potentials (RegPotential score) [42] for positions in the functional and non-functional STAT3 TFBSs detected by STAT-Scanner using the reference set of 22 genes (Table S1). For convenience, we considered a TFBS functional if it was supported by experimental STAT3 binding data; otherwise, the TFBS was considered non-functional. The distribution of PhastCons scores for the non-functional STAT3 TFBSs were skewed towards zero, while PhastCons scores for about 50% of the functional STAT3 TFBS exceeded 0.1 (Figure 3A). In contrast, the distribution of RegPotential scores, which measure the similarity of patterns to those in the known regulatory elements, was similar for positions of the functional and non-functional STAT3 TFBSs (Figure 3B). Next, we investigated the methylation-resistant CpG island features of the STAT3 TFBS-containing regions. Over-representation of the binding sequences for specific transcription factors, such as zinc-finger proteins, in CpG islands has been previously reported [43]. Most of the predicted STAT3 TFBSs are located inside CpG islands [44], but the genomic distribution is not significantly altered among the functional and non-functional STAT3 TFBSs (Figure 3C). Repeat elements [45] in the genomic sequence might compromise the functions of transcription factors, as none of the functional STAT3 TFBSs have been identified inside the repeated regions (Figure 3D). In summary, motif conservation, a major constraint that distinguishes between functional and non-functional STAT3 TFBSs, has therefore been included in STAT-Finder.

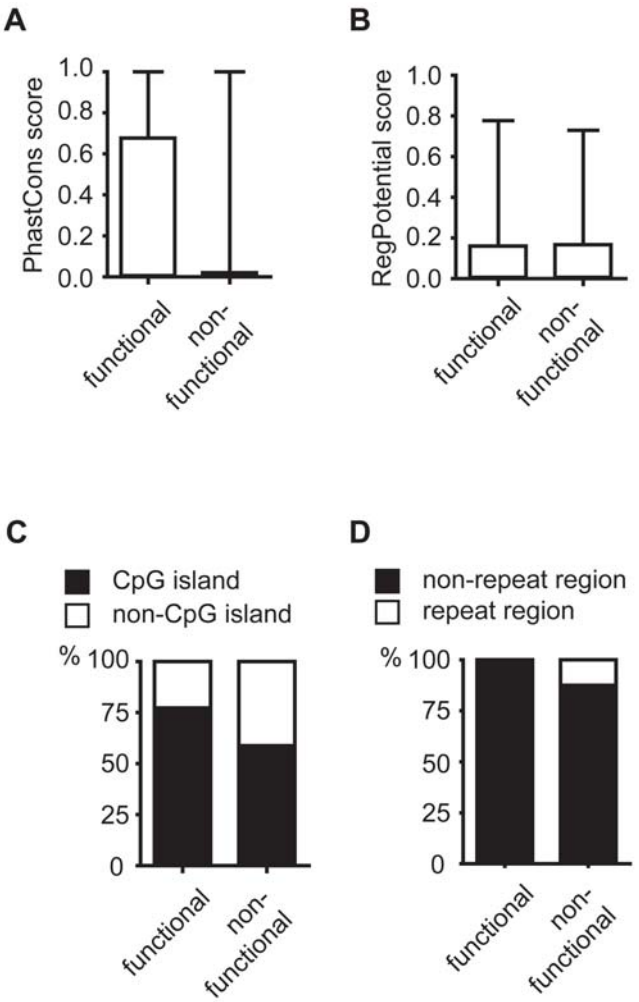

Figure 3. Score distribution of the functional vs. non-functional STAT3 TFBSs as predicted by STAT-Scanner. (A) PhastCons score, (B) Regulatory Potential score, (C) Percentage in the CpG island, and (D) Percentage in the Repeat region.
doi:10.1371/journal.pone.0006911.g003

Validation of STAT-Finder

We next evaluated the performance of STAT-Finder compared to other comparative methods, namely, EEL [46] and CONREAL [12]. Given that EEL performs pair-wise alignment based on the matches to a single PWM, we compared the performance of EEL using each PWM (V$STAT3_01 and V$STAT1_01) separately. Meanwhile, the performance of CONREAL was examined by combining both PWMs. We tested the prediction accuracy of STAT-Finder in the two positive data sets with STAT3 bindings. STAT-Finder exhibited better performance compared to EEL using V$STAT3_01, EEL using V$STAT1_01, or compared to CONREAL in predicting true STAT3 TFBSs in the 22 previously identified positive genes (Figure 4A). Note that both EEL and CONREAL failed to detect about 40–60% of true positive STAT3 sites even at the minimum cut-off value, while STAT-Finder found all of these. These data indicate that STAT-Finder showed better performance in terms of finding true positive STAT3 TFBSs that the other comparative programs missed. It was made more evident when we searched STAT3 TFBSs using EEL or CONREAL in the data sets with genome-wide STAT3 binding. Although the overall performance of the STAT-Finder was similar to EEL in detecting 56% of true STAT3 TFBSs, only STAT-Finder was capable of detecting the remaining 30% of the true sites (Figure 4B). Our data suggest that the improved sensitivity of STAT-Finder could be attributed to the usage of combined STAT-related PWMs, which evidently overcame the performance limitations of V$STAT3_01.

We next attempted genome-wide prediction of STAT3 binding in the human promoter regions. For this purpose, we first estimated the cut-off value of the motif conservation score (MCS) to identify conserved functional STAT3 TFBSs. The degree of conservation of the predicted TFBS, which was determined by calculating MCS, was integrated with the affinity scores by STAT-Scanner (See Methods). The confidence score at each MCS was evaluated using the 2.5 kb promoter sequences of all annotated human genes and orthologous mouse genes. The confidence score determines the probability that a given TFBS is not conserved by chance. As cut-off values of MCS increased, the total number of predicted STAT3 TFBSs decreased at a slower rate than the average number of aligned instances of control motifs, resulting in escalated confidence scores at MCS values higher than 0.9 (Figure

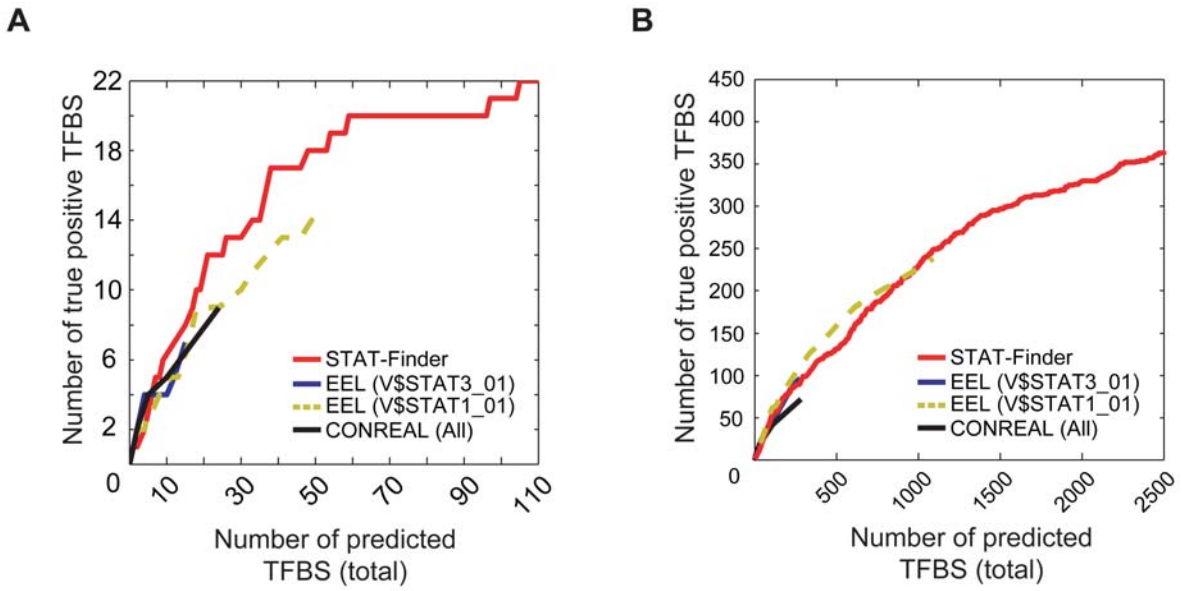

Figure 4. Performance comparison of the comparative alignment tools. Curves for the changes of the number of true binding sites detected using EEL (V$STAT3_01 or V$STAT1_01), CONREAL (All; combined PWMs of V$STAT3_01 and V$STAT1_01), or STAT-Finder, as a function of total number of predicted TFBSs (A) in the reference set of 22 genes (Table S1) and (B) in the genome-wide STAT3 ChIP-Seq dataset [39].
doi:10.1371/journal.pone.0006911.g004

S5). Using STAT-Finder, we performed a genome-wide search for STAT3 TFBSs in the human promoter regions. Among the 15461 human genes with identified orthologs in the mouse, about 7600 genes were predicted to have putative STAT3 binding sites within the 2.5 kb promoter region, at the probability threshold of 0.9. Significant enrichment of STAT3 TFBSs could be predicted at the proximal upstream regions of TSS using STAT-Scanner and STAT-Finder [35,39] (Figure S6).

Identification of novel STAT3 target genes in the cancer cells

Constitutive activation of STAT3 and over-expression of its target gene have been suggested to play critical roles in human carcinogenesis [12,31,47,48,49,50]. To determine whether or not STAT-Finder is useful in identifying novel STAT3 target genes, we applied this program to a group of genes that are over-expressed in human cancer cells. We integrated microarray data obtained from the expression module map of genes up-regulated in cancer [51] and data derived from the A549 cells over-expressing a constitutively active form of STAT3 [52].

Among the 33 genes that are commonly up-regulated, eleven have already been reported to be regulated by STAT3 (Table 1). Using this group of genes, we examined whether or not STAT-Finder could detect experimentally proven STAT3 TFBSs. It is noteworthy that we were able to analyze only a fraction of the promoter sequences, mainly due to alternative promoter usage and the poorly annotated TSS information available. STAT-Finder detected three putative STAT3 binding sites in the *JUNB* promoter region including one site that has previously been reported to be a STAT3 binding site [53] (Figure 5A). Using three different cell lines derived from human cancer patients, we confirmed STAT3 binding to the *JUNB* promoter by chromatin immunoprecipitation (Figure 5B). STAT-Finder also successfully detected one STAT3 TFBS in the Nicotinamide N-methyltransferase (*NNMT*) promoter region, a recently identified STAT3 target gene [54] (Figure 5C, D). Interestingly, STAT-Finder was unable to detect known STAT3 TFBS in the *MYC* promoter region (Figure 5E), even though *MYC* has been reported to be a STAT3 target [55]. It has also been reported that STAT3 binding to the promoter region of the *MYC* gene requires a site that is different from the consensus STAT3 binding sequences, but is similar to E2F TFBS, indicating that, in this case, STAT3 binding depends on the presence of other transcription factors [55]. Using primer sets that detect known STAT3 binding sites in the *MYC* promoter, we were able to confirm its binding upon IL-6 stimulation in HepG2 cells (Figure 5F). These results suggest that STAT-Finder could efficiently detect binding sites for STAT3 only if their binding does not depend on the presence of other *cis* or *trans* factors.

We next examined whether or not we can identify novel target genes of STAT3 using STAT-Finder. For this purpose, we selected genes with conserved TSS (Table 1) and determined the presence of putative STAT3 TFBSs using STAT-Finder in their promoter regions. STAT-Finder successfully detected putative STAT3 TFBSs with high probabilities in the promoter regions of *AKAP12* (A-kinase anchoring protein 12), *HIC2* (hyper-methylated in cancer 2), and *THBS1* (Thrombospondin 1). STAT3 binding to these predicted sites was experimentally confirmed by ChIP assay (Figure 6A–F). To verify the specificity of STAT-Finder, we also assayed the binding of STAT3 to the sites that were not conserved, but were present in the promoters of human orthologous genes. In contrast to the conserved STAT3 TFBSs, we could not detect STAT3 binding to the non-conserved STAT3 TFBSs in human cancer cell lines (Figure 6G). STAT3 binding to other predicted STAT3 TFBSs present in the promoter regions of *ATF3* (activating transcription factor 3), *DUSP5* (dual specificity phosphatase 5), *SERPINE1* (serpin peptidase inhibitor, class E), *NP* (nucleoside phosphorylase), and *SLC2A3* (solute carrier family 2, facilitated glucose transporter, member 3) were also experimentally validated (Figure S7). Finally, we studied whether or not other computation tools such as EEL or CONREAL could also accurately detect STAT3 target sites that have been identified and validated in this study. Of 10 promoter sequences containing experimentally proven 10 STAT3 binding sites (Figure 5, 6 and

Table 1. Putative STAT3 target genes.

Gene	Entrez ID	[a]Fold Change (log_2)	[a]FDR	[b]Cancer Module #	[b]Cancer Module (*P*-value)	Reported STAT3 regulation	Reported STAT TFBS	[c]Reference	Remark for experiment
AKAP12	9590	3.955	0	3	<1e-14	-	-	-	Putative target
ATF3	467	5.885	0	197	0.002153	-	-	-	Putative target
CCL2	6347	7.205	0	3	<1e-14	+	+	[73]	
CITED2	10370	2.911	0	3	<1e-14	+	+	[74]	
CXCL2	2920	2.092	0	197	0.021719	+	-	[75]	
DDEF2	8853	2.184	0	98	1.05E-12	+	-	[76]	
DUSP5	1847	2.782	0	98	0.001713	-	-	-	Putative target
ETS2	2114	3.039	0	197	0.000899	-	-	-	
FOSL1	8061	4.032	0	98	7.13E-05	-	-	-	
HIC2	23119	2.817	0	17	0.007665	-	-	-	Putative target
JUN	3725	3.194	0	197	5.36E-05	+	-	[77]	
JUNB	3726	2.436	0	17	1.07E-05	+	+	[53]	Positive control
LDLR	3949	3.548	0	3	<1e-14	-	-	-	
LOXL2	4017	2.139	0	3	<1e-14	-	-	-	
MAFF	23764	5.273	0	3	<1e-14	+	+	[78]	
MAP2K3	5606	2.739	0	18	0.001053	-	-	-	
MYC	4609	2.053	0	126	<1e-14	+	+	[55]	Positive control
NNMT	4837	2.041	0	3	<1e-14	+	+	[54]	Positive control
NP	4860	2.043	0	3	<1e-14	-	-	-	Putative target
NPC1	4864	5.916	0	18	1.1E-06	-	-	-	
PLAUR	5329	2.89	0	3	2.53E-12	-	-	-	
PLEC1	5339	2.982	0	18	0.016729	-	-	-	
PLEKHC1	10979	2.188	0	3	0.045975	-	-	-	
PMAIP1	5366	2.031	0	54	1.7E-05	+	-	[79]	
PXN	5829	2.217	0.098	18	5.01E-06	-	-	-	
SERPINE1	5054	2.312	0	3	<1e-14	-	-	-	Putative target
SGK	6446	2.826	0	3	0.004895	+	+	[80]	
SLC2A3	6515	6.191	0	17	2E-06	-	-	-	Putative target
TAF1A	9015	2.661	0.042	124	0.005465	-	-	-	
THBS1	7057	3.25	0	3	<1e-14	-	-	-	Putative target
UGCG	7357	6.265	0	3	2.21E-08	-	-	-	
WEE1	7465	2.172	0.069	57	<1e-14	-	-	-	
ZYX	7791	2.124	0	3	<1e-14	-	-	-	

[a]Analyzed microarray data of A549 cell line over expressing STAT3C [52] using SBEAMS [72].
[b]Analyzed data of the cluster in the Cancer Module Map [51] (http://robotics.stanford.edu/~erans/cancer/).
doi:10.1371/journal.pone.0006911.t001

S7), STAT-Finder predicted a total of 29 STAT3 binding sites including all of the 10 experimentally validated STAT3 binding sites. Meanwhile, EEL and CONREAL detected only 5 (50%) and 2 (20%) validated STAT3 binding sites among 23 and 6 total predictions, respectively, thereby indicating that STAT-Finder has better performance in terms of identifying novel target genes of STAT3 (Figure S8).

Discussion

We presented a computational framework for identifying functional STAT3 TFBSs in mammalian promoters. The first compartment, STAT-Scanner, was designed to predict functional STAT3 TFBSs with improved sensitivity. By using comparative motif-based alignments, STAT-Scanner was linked to STAT-Finder to minimize false positive predictions. Our proposed method was tested using previously identified STAT3 target genes and was successfully applied to the identification of novel target genes.

Our strategy in developing STAT-Finder relied on several assumptions. First, the DNA binding specificity of STAT3 is shared by other STAT family members. STAT transcription factors bind to similar DNA sequences, and the similar DNA binding specificity of various STAT transcription factors, such as STAT1, STAT5A/5B, or STAT6, have been experimentally proven [56]. It has also been noted that integration of the overlapping matches detected by matrices from the same family members greatly reduces the number of total predicted TFBSs, and hence decreases the rate of false positive detection [57]. Furthermore, it has been recently reported that roughly half of

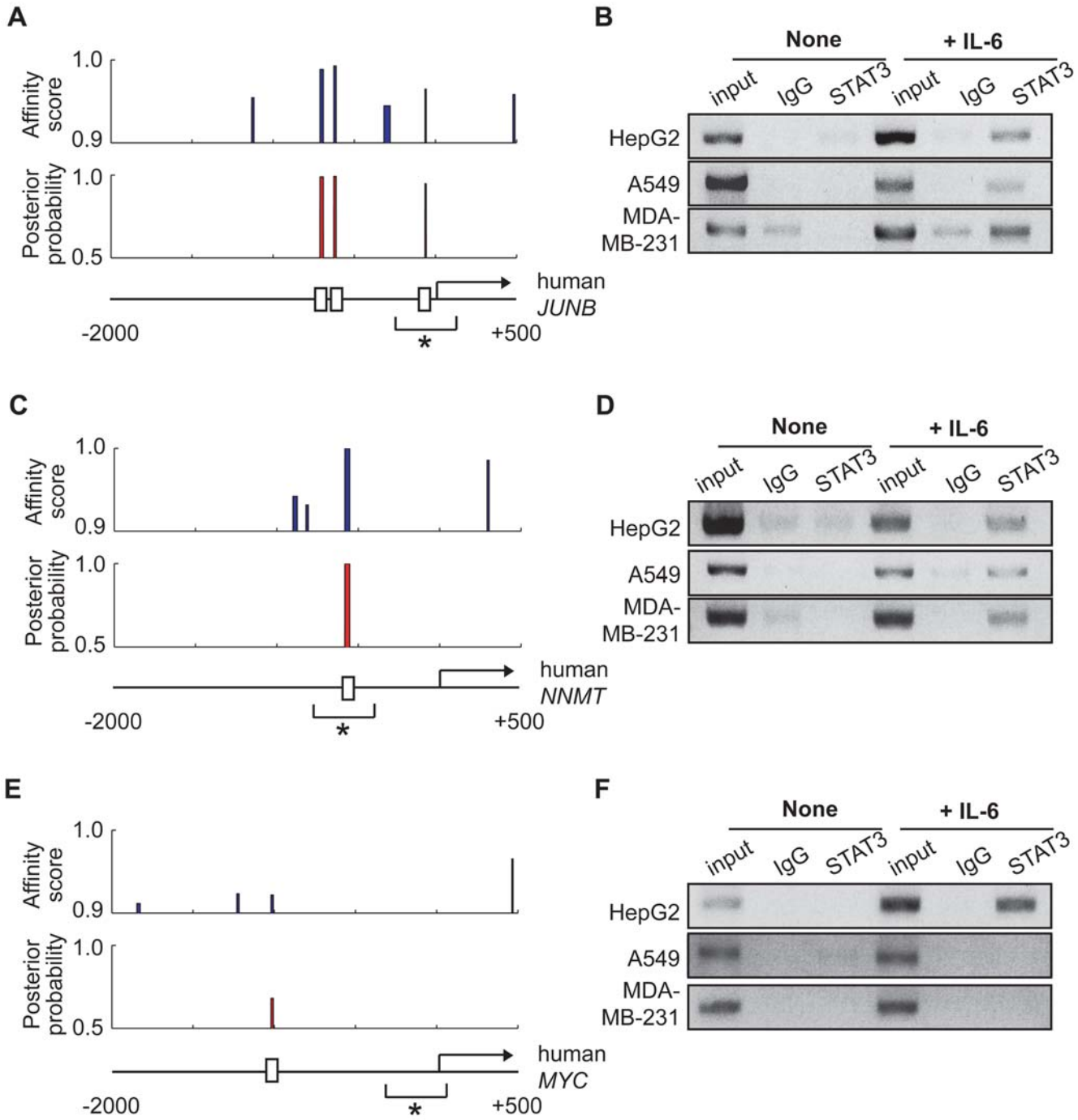

Figure 5. Experimental validation of STAT3 binding to the known STAT3 TFBSs. (A,C,E) The affinity score from STAT-Scanner (top) and the posterior probability from STAT-Finder (middle) of predicted STAT3 are plotted in the sliding windows for a 2.5-kb promoter region across the *JUNB* (A), *NNMT* (C), and *MYC* (E) genomic loci. The open square at bottom indicates the predicted TFBS with the posterior probability higher than 0.95; while the asterisk (*) in the promoter region depicts the known STAT3 TFBS. (B, D, F) Chromatin immunoprecipitation analysis with an anti-STAT3 antibody: Reported STAT3 TFBSs of *JUNB* (B), *NNMT* (D), and *MYC* (F) were PCR amplified using the primers specific binding sites (*) from the input and immunoprecipitated cell lysates, derived from the non-stimulated or IL-6 (10 ng/ml) + IL-6sR (10 ng/ml)-stimulated HepG2, A549, and MDA-MB-231 cells.

doi:10.1371/journal.pone.0006911.g005

TFs recognize multiple sequence motifs [58]. Therefore, a conventional motif scanning approach using a single PWM for each TF has an intrinsic limitation in detecting all functional TFBSs. As a result, the predictive power of STAT-Scanner was significantly enhanced by integrating STAT-related PWMs. The second assumption, used in the motif-based alignments, is that the relative locations of functional TFBSs are conserved among closely related mammalian species. In yeast, highly conserved TFBSs for a set of TFs exhibit relatively low spatial deviations (~150–200 bp) [20]. Likewise, we found that, for six mammalian species, known STAT3 TFBSs are located within a similar spatial distribution on each promoter.

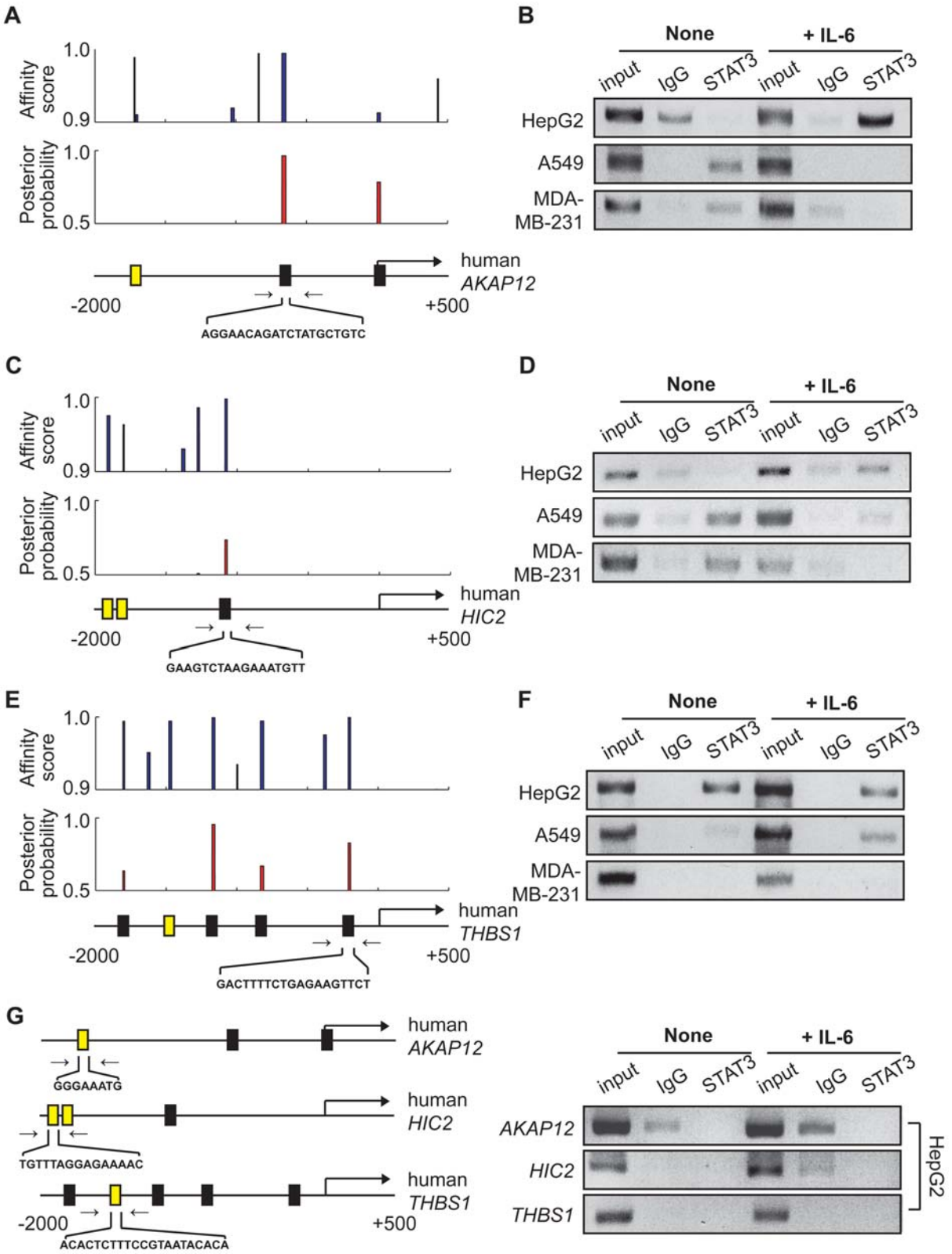

Figure 6. Experimental validation of STAT3 binding to the novel STAT3 TFBSs. (A, C, E) The affinity score (top, STAT-Scanner) and posterior probability (middle, STAT-Finder) of predicted STAT3 TFBSs are plotted in the sliding windows for a 2.5-kb promoter region across the *AKAP12* (A), *HIC2* (C), and *THBS1* (E) genomic locus. The closed square at the bottom indicates the predicted TFBS with posterior probability >0.5; while the yellow square shows the predicted TFBS with no conservation. (B, D, F) ChIP analysis with an anti-STAT3 antibody. Putative STAT3 TFBSs of the *AKAP12* (B), *HIC2* (D), and *THBS1* were PCR amplified using the primer sets indicated by inverse arrows. (G) ChIP analysis with an anti-STAT3 antibody. Predicted TFBSs with no conservation in the human *AKAP12*, *HIC2*, and *THBS1* genes were PCR amplified using the primer sets indicated by inverse arrows.
doi:10.1371/journal.pone.0006911.g006

Using STAT-Finder, we have identified a list of STAT3 target genes that are over-expressed in human cancer cells. Likewise, STAT3 binding to the predicted TFBSs has been experimentally verified in IL-6 stimulated human cancer cell lines. Interestingly, STAT3 was recruited to the predicted TFBS in a cell type-specific manner. For example, STAT3 binding to the predicted TFBSs in the promoter regions of the *AKAP12* and *HIC2* genes was observed in un-stimulated but not in IL-6 stimulated A549 and MDA-MB-231 cells. However, in the HepG2 cells, STAT3 was recruited to the same TFBS only after IL-6 stimulation (Figure 6). In contrast, STAT3 binding to the promoter regions of *MYC*, *SERPINE1*, *NP*, and *SLC2A3* was only detectable in IL-6 stimulated HepG2 cells, but not in A549 or MDA-MB-231 cells (Figure 6, Figure S7). Furthermore, it is evident that STAT3 binding to the predicted TFBSs in the promoters of the candidate target genes does not guarantee the expression of that gene. Although the expression of most of the target genes had been altered upon STAT3 binding to the promoter, we found that STAT3 binding to target sites did not always correlate with gene expression in the cell lines tested (Oh, YM, unpublished data). This suggests that STAT3 binding to target sites is not sufficient in inducing gene expression, and tissue-specific transcription factors, or trans-activators that specifying modification in the chromatin region may also be required [59,60,61,62].

A *cis*-regulatory module comprises a cluster of multiple TFBSs that cooperatively-interact with TFs to control gene expression. The identification of *cis*-regulatory modules for specific gene regulation is a challenging step towards understanding genome-wide transcription regulatory networks in mammalian genomes. Therefore, it is necessary to efficiently predict functional TFBSs for individual TFs. We expect that our comparative approach can be applied to other TFs with some restrictions. First, the efficiency of our program depends on the degree of evolutionary conservation among the six mammalian species. Therefore, DNA binding sites for TFs engaged in species-specific gene regulation may not be predicted. It is noteworthy that the frequent gain or loss of TFBSs in the intergenic regions leads to the evolution of transcriptional circuits [63]. Second, our program may not be applied to TFs that rely on other DNA binding proteins for recruitment into DNA. Third, because we only compared 2 kb of upstream promoter sequence relative to the annotated TSS, DNA binding sites of TFs that are enriched in regions distal to the TSS might be overlooked by our program. Although *cis*-regulatory regions that lie >100 kb away from the TSS exist, it has been suggested that most functional TFBSs are highly enriched in regions proximal to the TSS [40,64]. Another limitation is the amount and quality of annotated TSS information obtained from diverse mammalian species. With the exception of those from humans and mice, annotated TSS information for most of the mammalian genomes is not available, and correct TSS information is crucial for the identification of evolutionarily conserved TFBSs based on motif-based alignments. Obtaining accurate and reliable prediction of functional TFBSs in the promoter region is a critical step in deciphering the regulatory code of the complex transcription regulatory networks that govern diverse biological responses. Given that our proposed method is based on a multiple-motif model, we believe it can be applied to other TFs, with some modifications, and may serve as a basic tool to discover important *cis*-regulatory features.

Materials and Methods

Clustering of STAT3-related PWMs

We used dynamic programming to find the optimal gapped alignment between two PWMs. We denote by $\Theta_1 \in \mathbb{R}^{W_1 \times 4}$, $\Theta_2 \in \mathbb{R}^{W_2 \times 4}$ the two PWMs of length W_1, W_2 over $\Sigma = \{\mathrm{A,C,G,T}\}$ to align, where $\Theta^T_{k,w}$ represents row w, each entry is non-negative and $\sum_{l=1}^{4} \Theta_{k,wl} = 1$. The optimal pair-wise alignment can be found by the following steps. We first construct a matrix $\mathbf{F} \in \mathbb{R}^{(W_1+1)\times(W_2+1)}$ whose $(p+1, q+1)$-element $\mathbf{F}(p+1, q+1)$ is the score of the optimal alignment between the sub-matrices $\Theta^T_{1,1:p}$ and $\Theta^T_{2,1:q}$. Initially, we set $\mathbf{F}(1,1) = 0$, $\mathbf{F}(p+1,1) = p\delta$ and $\mathbf{F}(1,q+1) = q\delta$ for all p,q, where δ is a gap penalty. We then build up the matrix $\mathbf{F}$ using the following recurrence:

$$F(p+1, q+1) = \max \begin{cases} F(p,q) + s\left(\Theta^T_{1,p}, \Theta^T_{2,q}\right), \\ F(p,q+1) + \delta, \\ F(p+1,q) + \delta \end{cases}$$

where $s\left(\Theta^T_{1,p}, \Theta^T_{2,q}\right)$ is the match score which is defined by the Kullback-Leibler divergence

$$s\left(\Theta^T_{1,p}, \Theta^T_{2,q}\right) = 2\exp\left(-\mathrm{KL}\left(\Theta^T_{1,p} | \Theta^T_{2,q}\right)\right) - \frac{1}{2}$$

where

$$\mathrm{KL}\left(\Theta^T_{1,p} | \Theta^T_{2,q}\right) = -\sum_{l=1}^{4} \Theta^T_{1,pl} \log\left(\frac{\Theta^T_{2,ql}}{\Theta^T_{1,pl}}\right).$$

To define the similarity between the two PWMs, we assessed the optimal alignment by calculating the *P*-value. The *P*-value of the observed alignment score was calculated by estimating the score distribution of 1000 randomly-permuted PWMs via the Gaussian distribution. In this study, the gap penalty δ was set to 0.5.

Prediction of putative STAT TFBSs: STAT-Scanner

We searched putative TFBSs of the STAT family in input promoter sequences and evaluated their binding affinity scores using STAT-Scanner. Given a set of position count matrices obtained from TRANSFAC 9.4 [32], we reconstructed STAT-related PWMs to compute the *P*-values of the match scores using a method for calculating the exact distribution of scores [65,66]. Briefly, our method consists of three steps. First, we transformed each position count matrix into the corresponding position frequency matrix (PFM) by adding position-dependent pseudo-counts [34]. This position specific regularization leaves the conserved positions of the matrix relatively unchanged. Second, from the regularized PFM, we reconstruct a position weight matrix (PWM) whose element is the log-odds score between the PFM and background model, defined by a zero-order Markov chain. Then, the match score is defined by the sum of the log-odds scores. To account for the effect of the uneven distribution of "GC" and "AT" content, we used six different background models that were constructed based on clusters of nucleotide compositional vectors of the whole mouse promoter sequences available from Ensembl [67]. For clustering, we used the *k*-means clustering algorithm. Finally, to determine statistically significant TFBSs, we calculated the exact distributions of the match scores under the background model assumption. From the distributions, we calculated the type-I sequence error probability $\alpha_n(s)$, which measures the probability that at least one site within a sequence of length n ($n = 500$ as proposed by [34]) has a match score larger than or equal to s, under the assumption that the sequence is generated from the background model. We then converted the match score, s, into the affinity score, t, defined by 1- $\alpha_n(s)$. This

conversion makes it easy to define a threshold, γ, of the affinity score within a statistical hypothesis testing framework. In addition, it is also plausible to directly compare the affinity scores of different STAT-related PWMs.

Given a set of STAT-related PWMs and an input promoter sequence, STAT-Scanner first computes the nucleotide composition of the input sequence in order to select the nearest background model, and searches TFBSs whose affinity scores are larger than the threshold, with the PWMs constructed by the chosen background model. We used eight STAT-related PWMs, V$STAT_01, V$STAT1_01, V$STAT3_01, V$STAT5A_01, V$STAT5B_01, V$STAT5A_02, V$STAT1_02, and V$STAT_Q6, and combined all overlapping sites of the eight PWMs into one with the maximum affinity score. Notably, the max operator is applicable because the affinity scores of different PWMs are directly comparable.

Prediction of the conserved STAT TFBSs: STAT-Finder

STAT-Finder was designed to minimize the false positive discovery of predicted TFBSs using comparative sequence comparisons. It searches conserved sites within the promoters of six orthologous species by sequences of affinity scores. To diminish probabilities of misalignments, we used score sequences defined by STAT-Scanner as a first approximation of the conserved regulatory regions. We regarded a region with nonzero affinity scores within the score sequence as the regulatory region. We focus on the regulatory regions of multiple alignments by ignoring the non-conserved regions. We progressively aligned the six score sequences obtained from orthologous promoter sequences, according to the phylogenetic tree of all six mammalian species, and evaluated the degree of conservation by calculating the Motif Conservation Score (MCS). The MCS value of an aligned TFBS ranges from 0 (non-conserved) to 1 (most conserved).

Motif-based alignment tool consists of two main parts to align multiple sequences with different affinity scores of STAT TFBSs. The first part is a pair-wise global alignment module that finds an optimal alignment between two sequences. We adapted a variant version of the Needleman-Wunsch algorithm [68], with a modification in the scoring function for the match between two affinity scores. The second part is a progressive alignment module that determines a multiple motif alignment among orthologous promoter sequences derived from six mammalian species. The basic concept of this approach is to sequentially perform pair-wise alignments between two sequences of affinity scores, between a sequence and a profile, or between two profiles, according to the phylogenetic tree of the six species. The profile is a set of aligned sequences with gaps. This progressive alignment efficiently aligns multiple sequences with reasonable accuracy. In the given multiple motif alignment, we computed the motif conservation score (MCS) at each aligned position by taking the average of the aligned affinity scores. The affinity score of the gap was set to zero and the "N" character was not considered when calculating the average.

STAT-Finder has a unique feature that detects not only conserved binding sites but also non-conserved ones with very strong binding signals to rectify the unavoidable alignment error. Before describing our probabilistic model, we first explain our notations to define our model. Among the 8 STAT3-related PWMs, we denote by $\Theta_k \in \mathbb{R}^{W\times 4}$ the k^{th} PWM of length W over Σ, where $\Theta_{k,w}^T$ represents row w, each entry is non-negative and $\sum_{l=1}^{4}\Theta_{k,wl}=1$. The background model $\theta_0 \in \mathbb{R}^4$, which describes frequencies over the alphabet within non-binding sites, is defined by a zero-order Markov chain. We assume the background model is known in advance and estimated from the whole mouse promoter sequences.

Suppose we have a promoter sequence S_i which is a string of length L over the alphabet Σ. In order to allow for multiple binding sites per sequence, we represent the sequence S_i as a set of overlapping subsequences $S_{ij}^W=(S_{ij},S_{i(j+1)},\ldots,S_{i(j+W-1)})$ of length W starting at position $j \in I_i^W$, where S_{ij} denotes the letter at position j and $I_i^W=\{1,\ldots,L-W+1\}$. Let us introduce a latent variable matrix $Z_i^k \in \mathbb{R}^{2\times|I_i^W|}$ in which the j^{th} column vector Z_{ij}^k is defined as a 2-dimensional binary random vector $\left[Z_{ij1}^k,Z_{ij2}^k\right]^T$ such that $Z_{ij}^k=[0,1]^T$ if a binding site of the k^{th} PWM starts at position $j \in I_i^W$. Otherwise, $Z_{ij}^k=[1,0]^T$.

Our probabilistic model has the following specification for defining the joint distribution. First, latent variables Z_{ij}^k indicating the starting positions of binding sites of the k^{th} PWM are governed by the probability $\pi=[\pi_1,\pi_2]^T$ such that $\pi_1,\pi_2\geq 0$ and $\pi_1+\pi_2=1$. The prior probability of Z_{ij}^k is specified by

$$P\left(Z_{ij}^k|\pi\right)=\prod_{m=1}^{2}\pi_m^{Z_{ijm}^k}.$$

For each latent position Z_{ij}^k, the probability distribution of the subsequence S_{ij}^W is given by

$$P\left(S_{ij}^W|Z_{ij}^k,\Theta_k,\theta_0\right)=P\left(S_{ij}^W|\theta_0\right)^{Z_{ij1}^k}P\left(S_{ij}^W|\Theta_k\right)^{Z_{ij2}^k}$$

where

$$P\left(S_{ij}^W|\Theta_k\right)=\prod_{w=1}^{W}\prod_{l=1}^{4}\Theta_{k,wl}^{1\left(l,S_{i(j+w-1)}\right)}$$

$$P\left(S_{ij}^W|\theta_0\right)=\prod_{w=1}^{W}\prod_{l=1}^{4}\theta_{0l}^{1\left(l,S_{i(j+w-1)}\right)}$$

where $1\left(l,S_{i(j+w-1)}\right)$ is an indicator function which is 1 if $S_{i(j+w-1)}=l$, and 0 otherwise.

The objective of probabilistic inference in our model is to calculate the posterior probability $P\left(Z_{ij}^k|S_{ij}^W,\Theta_k,\theta_0\right)$ because this probability evaluates the degree of being a true binding site of the subsequence using our prior knowledge and given data. The posterior probability can be obtained by using Bayes' theorem

$$P\left(Z_{ij2}^k|S_{ij}^W,\Theta_k,\theta_0\right)=\frac{P\left(S_{ij}^W|\Theta_k\right)\pi_2}{P\left(S_{ij}^W|\theta_0\right)\pi_1+P\left(S_{ij}^W|\Theta_k\right)\pi_2}.$$

The degree of conservation of each subsequence can be easily incorporated into our probabilistic framework by assigning relatively higher prior probability π on Z_{ij}^k than non-conserved one. In this work, we used the following settings: $\pi_2=0.0002$ for non-conserved subsequence (the expected number of binding sites is 1 when the promoter sequence is 2500 (double stranded)) and $\pi_2=0.01\times\text{MCS}$. The eight different posterior probabilities of the eight STAT3-related PWMs across the latent positions are integrated by taking the maximum value and the probability cutoff value was set to 0.5.

Motif-based pair-wise alignment

We denote two promoter sequences of lengths m_i and m_j by $t^i=t_1^it_2^i\cdots t_{m_i}^i$ and $t^j=t_1^jt_2^j\cdots t_{m_j}^j$, where t_k^i and t_l^j are the affinity scores of STAT3 TFBSs. If the score is smaller than a threshold γ,

we cut the score to 0. We also set the score to -1 if the corresponding site contains the ambiguous "N" characters. With this setting, the optimal pair-wise alignment between t^i and t^j can be found by dynamic programming. We first construct a matrix $\mathbf{F} \in \mathbb{R}^{(m_i+1)\times(m_j+1)}$ whose $(k+1,l+1)$-element $\mathbf{F}(k+1,l+1)$ is the score of the optimal alignment between the segments $t_1^i \cdots t_k^i$ and $t_1^j \cdots t_l^j$. For initialization, we set $\mathbf{F}(1,1)=0, \mathbf{F}(k+1,1)=k\delta$ and $\mathbf{F}(1,l+1)=l\delta$ for all $k$and$l$, where δ is a gap penalty. We then build up the matrix $\mathbf{F}$ using the following recurrence:

$$F(k+1,l+1) = \max \begin{cases} F(k,l) + s\left(t_k^i, t_l^j\right), \\ F(k,l+1) + \delta, \\ F(k+1,l) + \delta \end{cases}$$

where $s\left(t_k^i, t_l^j\right)$ is the sequence match score between t_k^i and t_l^j which is defined by

$$s\left(t_k^i, t_l^j\right) = \begin{cases} 0 & \text{if } t_k^i = -1 \text{ or } t_l^j = -1, \\ \ln \frac{1-\gamma}{\left|t_k^i - t_l^j\right|} & \text{if } \left|t_k^i - t_l^j\right| \neq 0, \\ 10 & \text{if } \left|t_k^i - t_l^j\right| = 0 \text{ and } t_k^i, t_l^j > 0, \\ 0.1 & \text{if } t_k^i, t_l^j = 0 \end{cases}$$

We construct the optimal alignment by tracing back the choices that result in the final value of $\mathbf{F}(m_i+1, m_j+1)$.

Motif-based profile alignment

We denote two profiles of lengths m and n by $p = p_1 p_2 \cdots p_m$ and $q = q_1 q_2 \cdots q_n$, where p_k and $q_l$are the aligned affinity scores of STAT3 TFBSs constructed from disjoint promoter sequences indexed by $\boldsymbol{I}$ and $\boldsymbol{J}$, respectively. The pair-wise profile alignment can be also found by dynamic programming. The profile match score s_p is defined by the average of the sequence match score:

$$s_p(p_k, q_l) = \frac{1}{|\boldsymbol{I}||\boldsymbol{J}|} \sum_{i \in \boldsymbol{I}} \sum_{j \in \boldsymbol{J}} s\left(t_k^i, t_l^j\right)$$

where the sequence match score s is slightly modified to deal with the gap "-" by setting $s\left(-, t_l^j\right) = s\left(t_k^i, -\right) = \delta$ and $s(-,-)=0$. In this study, the score threshold γ and gap penalty δ were set to 0.8 and -0.1, respectively.

Estimation of MCS confidence score

To generate randomly shuffled control motifs, we manually aligned STAT-related PWMs without gaps by looking up the core regions (TTCCNGGAA). We excluded V$STAT5A_02 (homo-tetramer) because it was not aligned with other PWMs. The operation for random permutation was then applied to the aligned PWMs to generate 100 control motifs. Based on the assumption that the control motifs should have occurrence rates similar to the real motif, we selected 42 control motifs that detect similar numbers of TFBSs in the reference data set ($\pm$15%). Among them, we chose 10 motifs that were most dissimilar to V$STAT_01, based on the inter-motif distance measure (cut-off: 0.25) [69]. The confidence level at each MCS was then calculated using the following equation: (the total number of TFBSs of the real motif - the average number of TFBSs of the control motifs)/the total number of TFBSs of the real motif. This value represents the fraction of the number of conserved TFBSs above the ones that occurred by chance.

Retrieving information for promoter sequence

Human and mouse promoter sequences ($-2000\sim +500$ bp of the annotated transcription start sites) were downloaded from Table Browser of the UCSC genome browser [70]. We used hg18 and mm9 for human and mouse genome UCSC version, respectively. Orthologous promoter sequences of chimpanzee, orangutan, and rhesus were obtained by blatting the 2.5-kb human promoter sequences into the UCSC genome browser of each species [70]. Rat promoter sequences were obtained by blatting the 2.5-kb mouse promoter sequences into the UCSC genome browser of rat. For each 2.5-kb promoter sequence, PhastCons scores, Regulatory potential scores, CpG island, and regions for repeated elements were also obtained through Table Browser of the UCSC genome browser.

Genome-wide STAT3 ChIP-Seq data set was obtained from [39]. In this data set, we first selected 461 genes with STAT3 binding peaks located in 2.5kb promoter regions, among which 412 genes have at least one site predicted by STAT-Scanner (cut-off: 0.2) that is overlapped with experimentally identified regions (within 150 bp of STAT3 binding peaks). We next defined true positive sites as those that are overlapped with the STAT3 binding regions and that match to the highest scoring site predicted by STAT-Scanner, as suggested by [71].

Microarray data analysis

Cancer module map information was downloaded from the web browser (http://ai.stanford.edu/~erans/cancer/). We used Cancer Module 3, 17, 18, 54, 57, 98, 124, 126, and 197, which contain commonly up-regulated genes across liver cancer, B lymphoma, grade3 breast cancer, and stimulated macrophages. Microarray CEL data files of STAT3-C over-expressed cells were obtained from Dr. E.B. Haura (University of South Florida, Tampa) [52]. Microarray data was analyzed using the SBEAMS program [72]. The data set was normalized by the global quantile scaling method (GC-RMA) and filtered to include differentially expressed genes with more than two fold change, with FDR <0.1 and P-value <0.01 (t-test).

Cell culture

Human hepatocarcinoma cell line, HepG2, was maintained in MEM supplemented with 10% FBS (Hyclone, Logan, UT) and 1% penicillin/streptomycin (Invitrogen, Carlsbad, CA). Human lung carcinoma cell line, A549, and breast cancer cell line, MDA-MB-231, were cultured in DMEM with 10% FBS and 1% penicillin/streptomycin. For IL-6 stimulation, cells were treated with rhIL-6 (10 ng/ml) and rhIL-6sR (10 ng/ml) (R&D Systems, Minneapolis, MN) for 15 minutes.

Chromatin Immunoprecipitation (ChIP)

ChIP assays were performed as described with minor changes [64]. Cells were fixed in 1% formaldehyde for 15 min, harvested in buffer A (0.25% Triton X-100, 10 mM EDTA, 0.5 mM EGTA, 10 mM HEPES [pH 6.5]), and then resuspended in buffer B (200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 10 mM HEPES [pH 6.5]). Cells were then lysed in lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl [pH 8.1] with proteinase inhibitors). Chromatin sonication was performed three times for 40 s at setting 5.0 using a Branson 250 sonicator with a microtip. Fragmented chromatin was immunoprecipitated with STAT3 antibodies (SC-482, SC-483; Santacruz Technology, CA, and

S21320; Transduction Laboratories, Lexington, KY) for 4 hrs. After reversal of the cross-links and DNA precipitation, enriched DNA was analyzed by PCR amplification with primers that flank the predicted STAT3 TFBSs (Table S2).

Supporting Information

Figure S1 PWM similarity clustering. (A) Total 565 vertebrate TRANSFAC PWMs were clustered by pair-wise similarity comparison with the Kullback-Leibler divergence. The number of PWM clusters at different similarity P-value cut-offs is plotted. (B) PWM cluster at 10–7 P-value of similarity was represented by Cytoscape [73]
Found at: doi:10.1371/journal.pone.0006911.s001 (1.55 MB TIF)

Figure S2 Quality scores of STAT-related PWMs and clustered STAT-related PWMs in the known STAT3 TFBSs. (A) Histogram of PWM quality score for all 565 vertebrate PWMs derived from TRANSFAC ver. 9.4. (B) Number of STAT3 binding sites detected by combined PWMs. Forty STAT3 TFBSs [35] were used as reference dataset. (C) PWM quality score of STAT-related PWMs.
Found at: doi:10.1371/journal.pone.0006911.s002 (0.96 MB TIF)

Figure S3 STAT3 TFBS prediction using MATCH and MotifLocator. Curves for the changes of the number of true positive TFBSs detected using MotifLocator (A) or MATCH (B) in the reference set of 22 STAT3 target genes. PWM: V$STAT3_01, V$STAT1_01, or combined PWMs of V$STAT3_01 and V$STAT1_01 (All).
Found at: doi:10.1371/journal.pone.0006911.s003 (0.67 MB TIF)

Figure S4 Comparison of the TFBS prediction programs using the genome-wide STAT3 binding. Curves for the changes of the number of true positive TFBSs detected using MATCH (A) or MotifLocator (B) in the genome-wide STAT3 ChIP-Seq dataset.
Found at: doi:10.1371/journal.pone.0006911.s004 (0.75 MB TIF)

Figure S5 Estimation of MCS confidence scores. The graph displays confidence scores (dotted line) and predicted numbers of conserved TFBSs (solid line) at each MCS cut-off value.
Found at: doi:10.1371/journal.pone.0006911.s005 (0.48 MB TIF)

Figure S6 Genome-wide distribution of predicted STAT3 TFBSs. Using 2.5-kb promoter sequences of all annotated human reference genes, predicted STAT3 TFBSs with STAT-Scanner (blue line at top, P-value <0.1) or STAT-Finder (blue line at bottom, posterior probability >0.5) were plotted. The red line (random) shows the distribution of predicted TFBSs in the randomly permutated promoter sequences.
Found at: doi:10.1371/journal.pone.0006911.s006 (0.80 MB TIF)

Figure S7 Experimental validation of STAT3 binding to the novel STAT3 TFBS. The affinity score (top, STAT-Scanner) and posterior probability (middle, STAT-Finder) of the predicted STAT3 TFBS are plotted in the sliding windows for a 2.5-kb promoter region across the ATF3 (A), DUSP5 (C), SERPINE1 (E), NP (G), SLC2A3 (I), and CCL2 (K) genomic loci. The closed square at bottom indicates predicted STAT3 TFBS with posterior probability >0.5. (B, D, F, H, J, L) ChIP analysis with an anti-STAT3 antibody.
Found at: doi:10.1371/journal.pone.0006911.s007 (7.45 MB TIF)

Figure S8 Performance comparison of the comparative alignment tools for the STAT3 target genes identified in this study.
Found at: doi:10.1371/journal.pone.0006911.s008 (0.36 MB TIF)

Table S1 Lists of the reference set for known STAT3 TFBSs.
Found at: doi:10.1371/journal.pone.0006911.s009 (0.17 MB DOC)

Table S2 The information for primer sets used in ChIP experiment.
Found at: doi:10.1371/journal.pone.0006911.s010 (0.04 MB DOC)

Acknowledgments

We thank Dr. E.B. Haura (University of South Florida, Tampa) for sharing microarray data of A549 over-expressing STAT3-C.

Author Contributions

Conceived and designed the experiments: YMO JKK JYY. Performed the experiments: YMO JKK YC. Analyzed the data: YMO JKK JYY. Contributed reagents/materials/analysis tools: SC. Wrote the paper: YMO JKK JYY.

References

1. Brivanlou AH, Darnell JE, Jr. (2002) Signal transduction and the control of gene expression. Science 295: 813–818.
2. Emerson BM (2002) Specificity of gene regulation. Cell 109: 267–270.
3. Spiegelman BM, Heinrich R (2004) Biological control through regulated transcriptional coactivators. Cell 119: 157–167.
4. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21: 33–37.
5. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166–176.
6. Della Gatta G, Bansal M, Ambesi-Impiombato A, Antonini D, Missero C, et al. (2008) Direct targets of the TRP63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. Genome Res 18: 939–948.
7. Elnitski L, Jin VX, Farnham PJ, Jones SJ (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16: 1455–1464.
8. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137–144.
9. Hannenhalli S (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. Bioinformatics 24: 1325–1331.
10. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16–23.
11. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276–287.
12. Berezikov E, Guryev V, Plasterk RH, Cuppen E (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. Genome Res 14: 170–178.
13. Chang LW, Nagarajan R, Magee JA, Milbrandt J, Stormo GD (2006) A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. Genome Res 16: 405–413.
14. Doniger SW, Huh J, Fay JC (2005) Identification of functional transcription factor binding sites using closely related Saccharomyces species. Genome Res 15: 701–709.
15. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301: 71–76.
16. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254.
17. Kheradpour P, Stark A, Roy S, Kellis M (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res 17: 1919–1931.
18. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE (2000) Human-mouse genome comparisons to locate regulatory sites. Nat Genet 26: 225–228.
19. Xie TX, Wei D, Liu M, Gao AC, Ali-Osman F, et al. (2004) Stat3 activation regulates the expression of matrix metalloproteinase-2 and tumor invasion and metastasis. Oncogene 23: 3550–3560.
20. Pritsker M, Liu YC, Beer MA, Tavazoie S (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation. Genome Res 14: 99–108.
21. Leonard WJ, O'Shea JJ (1998) Jaks and STATs: biological implications. Annu Rev Immunol 16: 293–322.

22. Ihle JN (2001) The Stat family in cytokine signaling. Curr Opin Cell Biol 13: 211–217.
23. Lerner L, Henriksen MA, Zhang X, Darnell JE, Jr. (2003) STAT3-dependent enhanceosome assembly and disassembly: synergy with GR for full transcriptional increase of the alpha 2-macroglobulin gene. Genes Dev 17: 2564–2577.
24. Schaefer TS, Sanders LK, Nathans D (1995) Cooperative transcriptional activity of Jun and Stat3 beta, a short form of Stat3. Proc Natl Acad Sci U S A 92: 9097–9101.
25. Yoo JY, Wang W, Desiderio S, Nathans D (2001) Synergistic activity of STAT3 and c-Jun at a specific array of DNA elements in the alpha 2-macroglobulin promoter. J Biol Chem 276: 26421–26429.
26. Alonzi T, Maritano D, Gorgoni B, Rizzuto G, Libert C, et al. (2001) Essential role of STAT3 in the control of the acute-phase response as revealed by inducible gene inactivation [correction of activation] in the liver. Mol Cell Biol 21: 1621–1632.
27. Murray PJ (2007) The JAK-STAT signaling pathway: input and output integration. J Immunol 178: 2623–2629.
28. Takeda K, Clausen BE, Kaisho T, Tsujimura T, Terada N, et al. (1999) Enhanced Th1 activity and development of chronic enterocolitis in mice devoid of Stat3 in macrophages and neutrophils. Immunity 10: 39–49.
29. Yoo JY, Huso DL, Nathans D, Desiderio S (2002) Specific ablation of Stat3beta distorts the pattern of Stat3-responsive gene expression and impairs recovery from endotoxic shock. Cell 108: 331–344.
30. Bromberg JF, Wrzeszczynska MH, Devgan G, Zhao Y, Pestell RG, et al. (1999) Stat3 as an oncogene. Cell 98: 295–303.
31. Bromberg J (2002) Stat proteins and oncogenesis. J Clin Invest 109: 1139–1142.
32. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.
33. Jensen ST, Liu JS (2008) Bayesian Clustering of Transcription Factor Binding Motifs. Journal of the American Statistical Association 103: 188–200.
34. Rahmann S, Muller T, Vingron M (2003) On the power of profiles for transcription factor binding site detection. Stat Appl Genet Mol Biol 2: Article7.
35. Vallania F, Schiavone D, Dewilde S, Pupo E, Garbay S, et al. (2009) Genome-wide discovery of functional transcription factor binding sites by comparative genomics: The case of Stat3. Proc Natl Acad Sci U S A 106: 5117–5122.
36. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, et al. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31: 3576–3579.
37. Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, et al. (2002) INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. Bioinformatics 18: 331–332.
38. Jiang C, Xuan Z, Zhao F, Zhang MQ (2007) TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res 35: D137–140.
39. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106–1117.
40. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434: 338–345.
41. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
42. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, et al. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15: 1051–1060.
43. Fan S, Fang F, Zhang X, Zhang MQ (2007) Putative zinc finger protein binding sites are over-represented in the boundaries of methylation-resistant CpG islands in the human genome. PLoS ONE 2: e1184.
44. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261–282.
45. Smit A, Hubley R, Green P (1996–2007) RepeatMasker Open-3.0. http://wwwrepeatmaskerorg.
46. Palin K, Taipale J, Ukkonen E (2006) Locating potential enhancer elements by comparative genomics using the EEL software. Nat Protoc 1: 368–374.
47. Grandis JR, Drenning SD, Zeng Q, Watkins SC, Melhem MF, et al. (2000) Constitutive activation of Stat3 signaling abrogates apoptosis in squamous cell carcinogenesis in vivo. Proc Natl Acad Sci U S A 97: 4227–4232.
48. Hodge DR, Hurt EM, Farrar WL (2005) The role of IL-6 and STAT3 in inflammation and cancer. Eur J Cancer 41: 2502–2512.
49. Song JI, Grandis JR (2000) STAT signaling in head and neck cancer. Oncogene 19: 2489–2495.
50. Spano JP, Milano G, Rixe C, Fagard R (2006) JAK/STAT signalling pathway in colorectal cancer: a new biological target with therapeutic implications. Eur J Cancer 42: 2668–2670.
51. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090–1098.
52. Dauer DJ, Ferraro B, Song L, Yu B, Mora L, et al. (2005) Stat3 regulates genes common to both wound healing and cancer. Oncogene 24: 3397–3408.
53. Coffer P, Lutticken C, van Puijenbroek A, Klop-de Jonge M, Horn F, et al. (1995) Transcriptional regulation of the junB promoter: analysis of STAT-mediated signal transduction. Oncogene 10: 985–994.
54. Tomida M, Ohtake H, Yokota T, Kobayashi Y, Kurosumi M (2008) Stat3 up-regulates expression of nicotinamide N-methyltransferase in human cancer cells. J Cancer Res Clin Oncol 134: 551–559.
55. Kiuchi N, Nakajima K, Ichiba M, Fukada T, Narimatsu M, et al. (1999) STAT3 is required for the gp130-mediated full activation of the c-myc gene. J Exp Med 189: 63–73.
56. Ehret GB, Reichenbach P, Schindler U, Horvath CM, Fritz S, et al. (2001) DNA binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites. J Biol Chem 276: 6675–6688.
57. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21: 2933–2942.
58. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. Science 324: 1720–1723.
59. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, et al. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell 122: 33–43.
60. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39: 311–318.
61. Kiryu-Seo S, Kato R, Ogawa T, Nakagomi S, Nagata K, et al. (2008) Neuronal injury-inducible gene is synergistically regulated by ATF3, c-Jun, and STAT3 through the interaction with Sp1 in damaged neurons. J Biol Chem 283: 6988–6996.
62. Urnov FD (2003) Chromatin remodeling as a guide to transcriptional regulatory networks in mammals. J Cell Biochem 88: 684–694.
63. Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol 3: e99.
64. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20: 1377–1419.
65. Claverie JM, Audic S (1996) The statistical significance of nucleotide position-weight matrix matches. Comput Appl Biosci 12: 431–439.
66. Staden R (1989) Methods for calculating the probabilities of finding patterns in sequences. Comput Appl Biosci 5: 89–96.
67. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, et al. (2004) An overview of Ensembl. Genome Res 14: 925–928.
68. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443–453.
69. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104.
70. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.
71. Xie X, Rigor P, Baldi P (2009) MotifMap: a human genome-wide map of candidate regulatory motif sites. Bioinformatics 25: 167–174.
72. Marzolf B, Deutsch EW, Moss P, Campbell D, Johnson MH, et al. (2006) SBEAMS-Microarray: database software supporting genomic expression analyses for systems biology. BMC Bioinformatics 7: 286.
73. Liu X, Das AM, Seideman J, Griswold D, Afuh CN, et al. (2007) The CC chemokine ligand 2 (CCL2) mediates fibroblast survival through IL-6. Am J Respir Cell Mol Biol 37: 121–128.
74. Leung MK, Jones T, Michels CL, Livingston DM, Bhattacharya S (1999) Molecular cloning and chromosomal localization of the human CITED2 gene encoding p35srj/Mrg1. Genomics 61: 307–313.
75. Hartner A, Goppelt-Struebe M, Hocke GM, Sterzel RB (1997) Differential regulation of chemokines by leukemia inhibitory factor, interleukin-6 and oncostatin M. Kidney Int 51: 1754–1760.
76. Snyder M, Huang XY, Zhang JJ (2008) Identification of novel direct Stat3 target genes for control of growth and differentiation. J Biol Chem 283: 3791–3798.
77. Sasaki A, Yasukawa H, Suzuki A, Kamizono S, Syoda T, et al. (1999) Cytokine-inducible SH2 protein-3 (CIS3/SOCS3) inhibits Janus tyrosine kinase by binding through the N-terminal kinase inhibitory region as well as SH2 domain. Genes Cells 4: 339–351.
78. Yang Y, Ochando J, Yopp A, Bromberg JS, Ding Y (2005) IL-6 plays a unique role in initiating c-Maf expression during early stage of CD4 T cell activation. J Immunol 174: 2720–2729.
79. Bai Y, Ahmad U, Wang Y, Li JH, Choy JC, et al. (2008) Interferon-gamma induces X-linked inhibitor of apoptosis-associated factor-1 and Noxa expression and potentiates human vascular smooth muscle cell apoptosis by STAT3 activation. J Biol Chem 283: 6832–6842.
80. Firestone GL, Giampaolo JR, O'Keeffe BA (2003) Stimulus-dependent regulation of serum and glucocorticoid inducible protein kinase (SGK) transcription, subcellular localization and enzymatic activity. Cell Physiol Biochem 13: 1–12.