BMC Genomics

BioMed Central

Research article

Open Access

Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression

Katsuhiko Murakami*, Toshio Kojima and Yoshiyuki Sakaki

Address: RIKEN Genomic Sciences Center, 1-7-22, Suehiro-cho, Tsurumi, Yokohama, Kanagawa, JAPAN

Email: Katsuhiko Murakami* - katsu@gsc.riken.go.jp; Toshio Kojima - tkojima@gsc.riken.go.jp; Yoshiyuki Sakaki - sakaki@gsc.riken.go.jp

* Corresponding author

Published: 23 February 2004

BMC Genomics 2004, **5**:16

Received: 19 May 2003

Accepted: 23 February 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/16

Abstract

Background: Gene expression is regulated mainly by transcription factors (TFs) that interact with regulatory *cis*-elements on DNA sequences. To identify functional regulatory elements, computer searching can predict TF binding sites (TFBS) using position weight matrices (PWMs) that represent positional base frequencies of collected experimentally determined TFBS. A disadvantage of this approach is the large output of results for genomic DNA. One strategy to identify genuine TFBS is to utilize local concentrations of predicted TFBS. It is unclear whether there is a general tendency for TFBS to cluster at promoter regions, although this is the case for certain TFBS. Also unclear is the identification of TFs that have TFBS concentrated in promoters and to what level this occurs. This study hopes to answer some of these questions.

Results: We developed the cluster score measure to evaluate the correlation between predicted TFBS clusters and promoter sequences for each PWM. Non-promoter sequences were used as a control. Using the cluster score, we identified a PWM group called PWM-PCP, in which TFBS clusters positively correlate with promoters, and another PWM group called PWM-NCP, in which TFBS clusters negatively correlate with promoters. The PWM-PCP group comprises 47% of the 199 vertebrate PWMs, while the PWM-NCP group occupied 11 percent. After reducing the effect of CpG islands (CGI) against the clusters using partial correlation coefficients among three properties (promoter, CGI and predicted TFBS cluster), we identified two PWM groups including those strongly correlated with CGI and those not correlated with CGI.

Conclusion: Not all PWMs predict TFBS correlated with human promoter sequences. Two main PWM groups were identified: (1) those that show TFBS clustered in promoters associated with CGI, and (2) those that show TFBS clustered in promoters independent of CGI. Assessment of PWM matches will allow more positive interpretation of TFBS in regulatory regions.

Background

Understanding the regulation of gene expression is a crucial issue in molecular biology. Since gene expression is mainly regulated by transcription factors (TFs), the elucidation of relationships among TFs, their binding sites (TFBS) and their controlling genes, is of great importance.

Although TFBS can be predicted by computer searches on DNA sequences, false positives (FP) are often produced. Several computer programs use position weight matrices (PWMs) [1] to predict TFBS *in silico*, including MatInspector [2], MATCH [3], and TFBS perl modules [4]. PWMs represent positional base preferences or frequencies constructed by a set of experimentally determined TFBS, and typically correspond to a single TF. The transcription factor database, known as TRANSFAC, is a widely used collection of PWMs [5]. TRANSFAC provides several PWMs for single TFs with different quality levels. Computer programs predict TFBS from DNA sequences, which are the same or similar to known TFBS. The low information contents in the matrices leads to many false positives, due to the weak preference or shortness of the site length (6–30 bp). Various strategies have been proposed to allow correct identification of true positives (TPs) from predicted TFBS. One approach is to employ information from conserved regions in DNA sequences between different species, known as phylogenetic footprinting. Bayes block aligner (BBA) is a tool used to extract conserved regions from an alignment of two DNA sequences [6]. It was demonstrated that it could identify binding sites of muscle-specific transcription factors [6]. Another approach is to identify multiple TFBS that form a structural cluster on a DNA sequence coordinate. This seems a reasonable technique because the density of predicted TFBS in promoter sequences is reported to be higher than non-promoter sequences, especially in the region 300 bp upstream from the transcription start site [7]. Genes are regulated by interactions with multiple functional TFs in metazoans [8]. Therefore, many promoter prediction programs, such as promoterscan [9], TSSG, and TSSW [10], have been developed based on the density of TFBS. The identification of genuine TFBS by searching clusters of predicted TFBS has been successful; however, these studies were evaluated with only specific genes and TF sets, such as those found in *Yeast*[11], *Drosophila* (early developmental enhancer) [12-14], liver [15], LSF and muscle specific regulatory regions [16,17]. It is unknown whether this method is applicable to other species, or genes. Although many vertebrate promoter sequences have CpG islands (CGI), the relationship between clusters of predicted TFBS and CGI is often underestimated [18]. Another strategy for the identification of putative TFBS includes a combinatorial approach that uses both phylogenetic footprinting and cluster analysis [12,15,19]. The program rVISTA utilizes information from conserved regions between human and mouse, in addition to clusters of TFBS predicted by the MATCH (BIOBASE) program [19]. This approach was evaluated using several known TFs (AP-1, NFAT, and GATA-3) and genes from the cytokine gene cluster. It remains unclear whether the properties used for clusters of TFBS are general and can be applied to other TFs or regulatory regions. Several reports have described methods for determining the statistical significance of predicted TFBS [11,12,17,20-22]. These studies assume the use of appropriate PWMs to identify clustered TFBS. To determine if a particular cluster is genuinely related to the promoter, it is important to assess clusters of predicted TFBS for each individual PWM. This is done using real non-promoter sequences for the appropriate selection of the PWM and for the interpretation of clusters of predicted TFBS. Most of these studies use specific sets of coregulated genes to identify common predicted TFBS clusters, and therefore cannot be applied directly to the study of general properties of promoters.

In this study, we developed a measure that evaluates the degree of concentration of predicted TFBS to clarify whether predicted TFBS have a tendency to cluster in human promoter sequences rather than in non-promoter sequences for each PWM. We identified some PWMs in which predicted TFBS clusters occur more significantly in promoter than non-promoter sequences and vice versa. Using partial correlations among three properties (promoters, CGI and clusters of predicted TFBS), we identified two PWM groups, (1) those in which TFBS cluster in promoters as a result of the presence of CpG islands, and (2) those in which TFBS cluster in promoters independent of CpG islands. We show that transcription factors corresponding to the latter PWM group tend to be tissue-specific. In summary, this analysis is useful for the interpretation of predicted TFBS in regulatory regions.

Results

Divergent preferences of TFBS for promoter sequences

We determined whether predicted TFBS formed clusters in human promoter sequences or in non-promoter sequences for each PWM using the cluster score described in the Method section. The higher the cluster score (derived from a logarithm of the p-value), the stronger the cluster of predicted TFBS is related to the promoter sequence. The threshold T, used to determine whether a cluster of predicted TFBS is found on a sequence, was calculated simultaneously. Since a prediction for the presence of a TFBS was performed for each PWM, an assessment for TFBS clusters was performed using the cluster score for each PWM. As a result, a number of PWMs do not tend to have clusters of TFBS in the promoter sequence. We observed a divergence of cluster scores. Of the 199 vertebrate PWMs in TRANSFAC, 94 (47%) PWMs had significantly high cluster scores, while 22 (11%) PWMs had significantly low cluster scores. The remaining 83 (42%) PWMs did not show significant cluster scores. A p-value of 1.0% was used to identify the above PWM set with Bonferroni correction for multiple testing ([23] Section 3.8). Figure 1 shows a histogram of cluster scores. Although these results were derived from genes on chromosome 20, the results from other chromosomes were

Table 1: Top 50 PWMs for chromosome 20 sorted by cluster score *S* in descending order. Each column represents rank number, accession number in TRANSFAC, identifier in TRANSFAC, cluster score, and threshold.

Rank	ACCESSION	ID	S	T
1	M00736	E2F1DP1_01	189.3	2.75
2	M00332	WHN_B	176.0	1.90
3	M00652	NRF1_Q6	122.0	0.93
4	M00649	MAZ_Q6	117.2	4.35
5	M00491	MAZR_01	111.4	1.78
6	M00739	E2F4DP2_01	103.8	0.93
7	M00737	E2F1DP2_01	103.6	0.94
8	M00108	NRF2_01	81.4	0.92
9	M00665	SP3_Q3	72.1	2.39
10	M00706	TFIII_Q6	61.4	4.23
11	M00740	E2F1DP1RB_01	58.4	0.90
12	M00324	MINI20_B	58.2	1.61
13	M00032	CETS1P54_01	57.3	3.70
14	M00743	CETS168_Q6	51.1	1.75
15	M00341	GABP_B	48.6	0.88
16	M00055	NMYC_01	41.1	0.90
17	M00329	PAX9_B	39.2	0.73
18	M00243	EGR1_01	37.3	0.87
19	M00072	CP2_01	36.5	1.66
20	M00054	NFKAPPAB_01	35.5	0.85
21	M00056	MYOGNF1_01	35.1	1.34
22	M00694	E4F1_Q6	35.0	0.86
23	M00738	E2F4DP1_01	34.9	0.91
24	M00143	PAX5_01	34.7	0.84
25	M00235	AHRARNT_01	34.6	0.92
26	M00698	HEB_Q6	33.6	0.91
27	M00039	CREB_01	33.6	1.00
28	M00514	ATF4_Q2	33.1	1.71
29	M00650	MTF1_Q4	31.4	0.88
30	M00194	NFKB_Q6	30.8	0.82
31	M00007	ELK1_01	30.0	0.85
32	M00733	SMAD4_Q6	29.7	0.81
33	M00261	OLF1_01	28.8	0.84
34	M00017	ATF_01	26.7	0.98
35	M00053	CREL_01	25.6	0.81
36	M00691	ATF1_Q6	25.5	0.89
37	M00244	NGFIC_01	25.2	0.88
38	M00041	CREBP1CJUN_01	24.9	1.00
39	M00086	IK1_01	24.2	0.90
40	M00287	NFY_01	24.0	1.95
41	M00466	HIF1_Q5	22.7	0.90
42	M00634	GCM_Q2	22.6	0.84
43	M00273	R_01	21.8	0.85
44	M00373	PAX4_01	21.7	2.57
45	M00097	PAX6_01	21.5	1.15
46	M00134	HNF4_01	21.1	0.64
47	M00670	TCF1P_Q6	21.1	0.80
48	M00057	COMP1_01	21.1	0.59
49	M00035	VMAF_01	21.0	1.32
50	M00222	HAND1E47_01	20.3	0.81

similar as described in the following subsection. PWMs with high cluster scores are shown in Table 1. Some of the PWMs have thresholds T (of accumulated score C) equal to or less than 1.0. This indicates that the occurrence of single predicted TFBS is more discriminative than clusters. Sequence logo [24] of the top three PWMs are depicted in

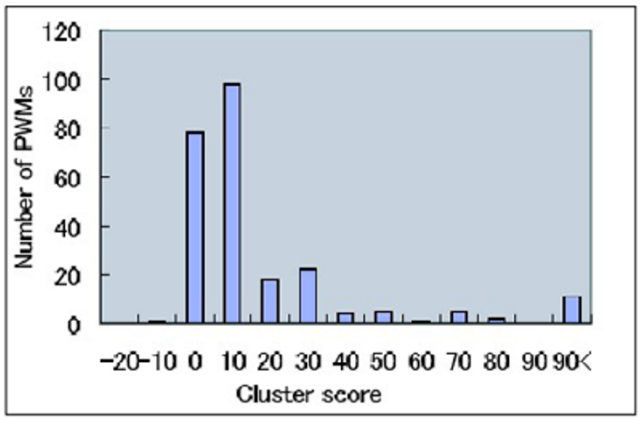

Figure 1
A histogram of cluster scores for PWMs. Each number of X-axis indicates the maximum score of PWMs in the bin.

Fig. 2-(a). See additional file 1 'PWMs sorted by cluster score' for the entire PWM list.

Cluster scores for different datasets

To assess the robustness of the cluster score we compared cluster scores for different datasets from chromosomes 20, 21 and 22, respectively. Fig. 3 shows the correlation of cluster scores between chromosomes 20 and 21 (a), and between chromosomes 20 and 22 (b). Some PWMs, the matches of which were detected on less than 50 subsequences, are not shown. The correlation coefficient points were 0.91 (a) and 0.93 (b).

Correlations among promoter sequences, CpG islands, and clusters

About half of the human coding genes have a compositional bias for CGI over transcription start sites [25]. It is possible that some of the predicted TFBS clusters might be the consequence of the existence of a CGI. To investigate this possibility, we computed partial correlation coefficients of three categories of promoters, CGI and predicted TFBS clusters. In general, a partial correlation coefficient measures the correlation between any pair of variables when other, specified variables, have been held constant. For example, a partial correlation coefficient $r_{IC.P}$ is the correlation between *I* and *C* while controlling for *P*, where *I* denotes CGI, *C* denotes accumulated score (strength of predicted TFBS clusters, see Methods) and *P* denotes promoters. If we calculate simple correlation coefficients, r_{PI} = 0.69, r_{IC} ranged from -0.25 to 0.57 for various PWMs, and r_{PC} ranged from -0.24 to 0.53 for various PWMs. These correlation coefficients are apparent ones. The partial correlation coefficients provide essential information and pure correlations, without the effect of the third variable. Fig. 4 shows a plot of $r_{IC.P}$ against $r_{PC.I}$ for various PWMs. For most of the PWMs, $r_{PC.I}$ is positive, although not particularly high (<0.3). This implies a correlation between clusters of these PWM matches and promoter sequences, separate to the effect of CGI. For the PWMs in the right circle in Fig. 4, $r_{PC.I}$ is high and $r_{IC.P}$ is approximately zero, where the cluster is more correlated with the promoter than the CGI. Some PWMs have a negative $r_{PC.I}$, implying the absence of promoter sequences for these PWM matches. For the PWMs in the top circle in Fig. 4, $r_{IC.P}$ is high and $r_{PC.I}$ is approximately zero, suggesting that the correlation between promoters and clusters for these PWM matches is attributable to the presence of the CGI. While these promoters and clusters do not correlate directly, they appear to correlate because both are associated with CGI.

Using these two values, we identified two PWM sets, (1) a CGI-related set (37 PWMs, Table 2) in which TFBS clusters are correlated with CGI (independent of promoter), and (2) a CGI-independent set (54 PWMs, Table 3) in which clusters of TFBS are correlated with promoters (independent of CGI). These sets were used for the following analysis.

Correlation between clusters of predicted TFBS and gene expression

Since all widely expressed, or housekeeping, genes have CGI [25], it is possible that clusters of PWM matches for CGI-independent sets are associated with tissue specific promoters. For this reason we examined the relationship between clusters of PWM matches and the tissue specificity of the associated genes using published gene expression data ([26,27]). The two resources used for this analysis are not consistent. Some genes annotated as housekeeping genes in one resource are referred to as tissue specific in another resource. We refer to these genes as mixed annotated genes. Genes with associated expression data were analysed and of these 72 were identified among the gene set covering the three chromosomes used in this study. They included 12 housekeeping genes, 9 mixed annotated genes, and 51 tissue specific genes. With the CGI-independent PWM sets we detected promoters with clusters of PWM matches. These clusters have significantly high Z-scores (see Methods) based on the accumulated score *C* in randomly generated DNA sequences (as control) with the same dinucleotide frequency of each promoter sequence. Table 4 shows the 40 genes detected, the *DCC* score (described below), their tissue specificity and start_p score. These genes are sorted according to the *DCC* score indicating the extent of association with CGI-independent PWMs over CGI-related PWMs. Results show that tissue specific genes tend to have high *DCC* scores and that transcription factors corresponding to CGI-independent PWMs are related to tissue specific genes. If we extract

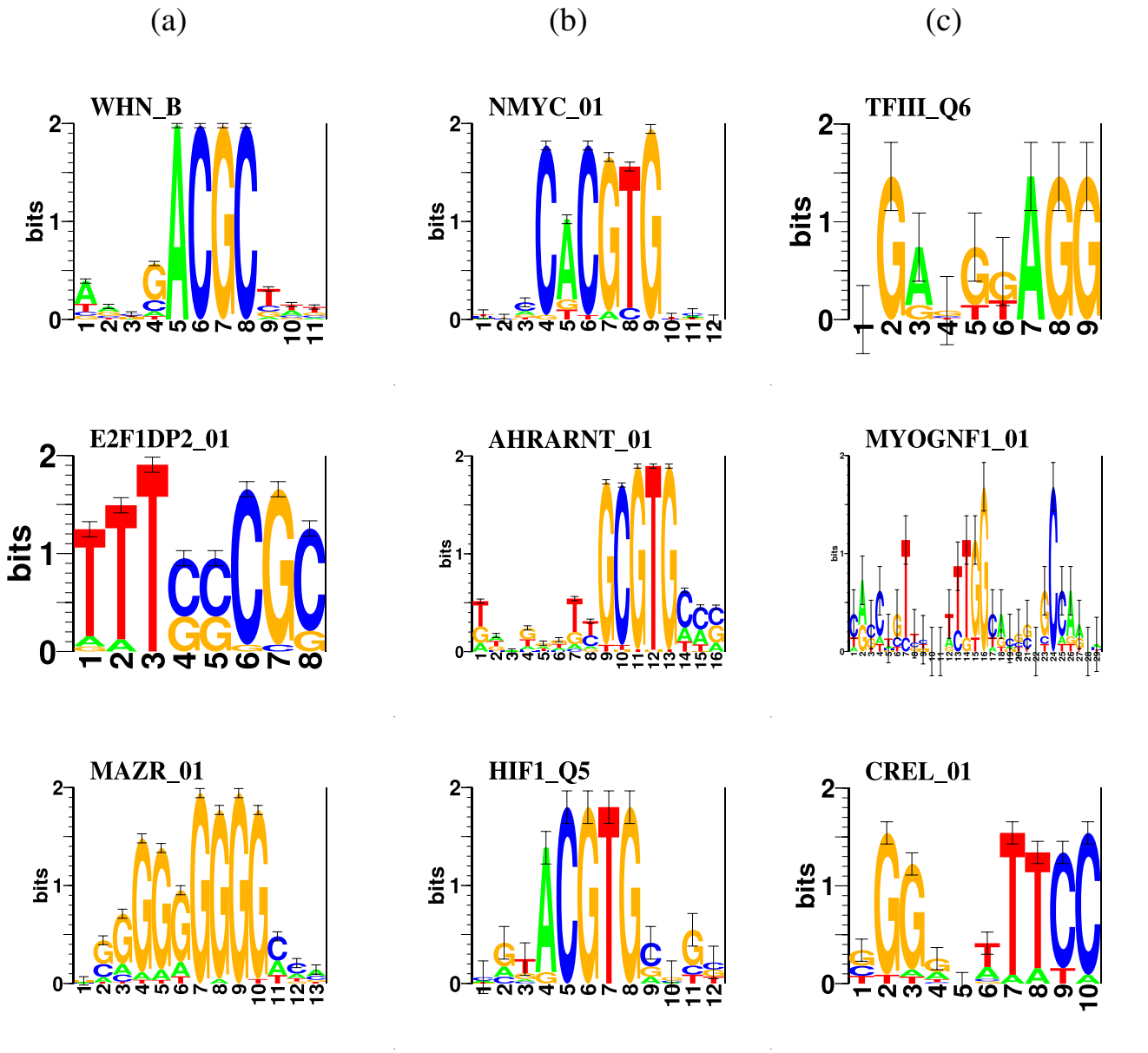

Figure 2
Sequence logos. (a) Top three PWMs from Table 1, (b) representative PWMs from Table 2, (c) representative PWMs from Table 3.

20 genes which have *DCC* scores equal to or higher than -0.03, 18 (90%) of these are tissue specific genes. The 72 genes with known gene expressions data included 51 tissue specific genes (71%). The p-value of the event of extraction was 0.04 under cumulative hypergeometric distribution. The p-value of the ranking of the two groups (11 housekeeping and 29 tissue-specific) in Table 4 was 0.01 by Wilcoxon rank test. Note that *DCC* is not correlated with the CGI score (start_p).

Discussion

Clusters of TFBS are an important property of regulatory regions [7,8,19,28]. To determine if this is a general tendency for PWM matches and all protein coding genes, we have developed a measure that evaluates the correlation between predicted TFBS concentrations and promoter sequences. We then examined the correlation for individual PWMs using an unbiased sequence set. Our results show that not all TFBS are clustered in promoter

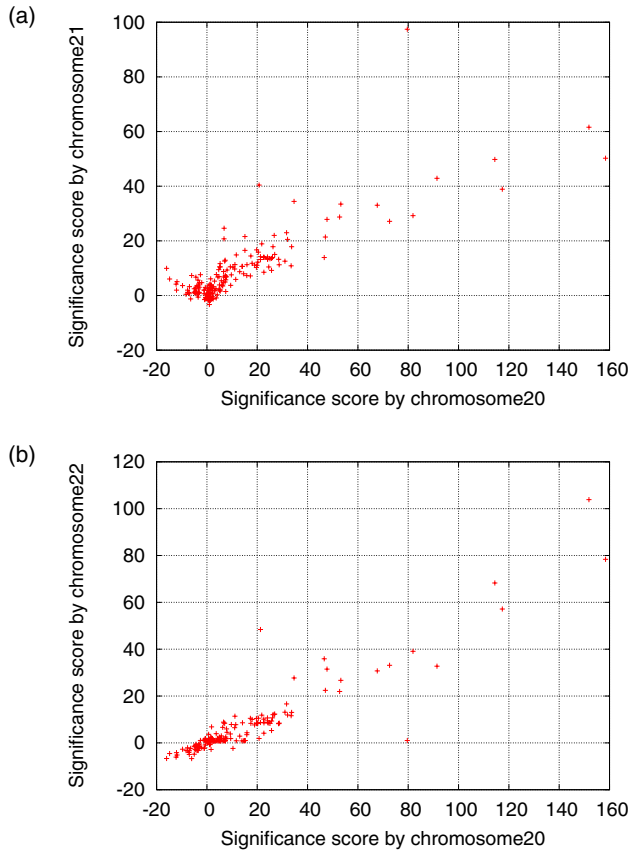

Figure 3
Title: Correlation of cluster scores (a) between chromosomes 20 and 21, (b) chromosomes 20 and 22.
Each dot represents a distinct PWM (defined by the TRANSFAC matrix). The correlation coefficients were (a) 0.91 and (b) 0.93.

sequences. We found that TFBS clusters corresponding to 47% of PWMs are positively correlated with promoter sequences, and that TFBS clusters corresponding to around 11% of PWMs are negatively correlated with promoter sequences.

It is important to ascertain the relationship between cluster scores of PWMs and CGI, because CGI are a prominent feature of promoter sequences. The consensus sequences of the top-ranked PWMs (Table 1) are, 'ANNGACGCTNN' (WHN_B), 'TTTCSCGC' (E2F1DP1_Q6), 'NSGGGGGGGMCN' (MAZR_01), and 'GGGGAGGG' (MAZ_Q6), where S represents C or G, M represents A or C, and N represents any bases. The sequence logos of the PWMs are depicted in Fig. 2-(a). The G+C % of base composition of each matrix is 70%, 56%, 91%, and 86%, respectively. The sequences with high cluster scores appear to be GC-rich. Larsen *et al* found that 57% of human genes are associated with CGI, that all housekeeping genes have CGI covering transcription start sites (TSS), and that 40% of tissue specific genes have CGI [25]. Therefore, the association of PWM-PCP with CGI may be significant, and CGI-related PWMs may play important roles in housekeeping regulation.

To evaluate the relationship between PWM-PCP and CGI, we calculated the partial correlation coefficient for each PWM. In general, if a correlation coefficient r_{XY} is not small and $r_{XY.Z}$ (defined in Methods) ≈ 0, the probable hypotheses concerning cause and effect will be either 1) the correlation of *X* and *Y* is a consequence of *Z*, or 2) *Z* intervenes between *X* and *Y*. For the PWMs in the top circle in Fig. 4, $r_{IC.P}$ is high and $r_{PC.I}$ is approximately zero. This suggests that the correlation between promoters and TFBS clusters is attributable to the presence of the CGI and that while they do not directly correlate they appear to because both independently correlate with CGI. The characteristic PWMs in Table 2 are NMYC_01 (M00055:0.25), AHRARNT_01 (M00235:0.23), and HIF1_Q5 (M00466:0.22), where parentheses include the accession number referred to in TRANSFAC and the recorded $r_{IC.P}$ (Y-value in Fig. 4). Sequence logos are depicted in Fig. 2-(b). The PWMs in the middle right circle in Fig. 4 have an $r_{IC.P}$ of approximately zero and a high $r_{PC.I}$ $r_{PC.I}$ value showing that the cluster is correlated with promoters independent of CGI. The predicted TFBS clusters corresponding to these PWMs could not be explained by the presence of CGI. Some of these PWMs have thresholds *T* less than 1.0 indicating that even the single occurrence of a predicted TFBS is more discriminative than clusters. Particular examples with high recorded $r_{PC.I}$ values and values for $r_{IC.P} < 0.1$ are TFIII_Q6 (M000706), MYOGNF1_01 (M00056) and CREL_01 (M00053). Sequence logos are depicted in Fig. 2-(c). TFIII_Q6 is a matrix associated with a general transcription factor II-I with the consensus sequence RGAGGKAGG, where the K represents G or T. The matrix TFIII_Q6 contains many 'G', and 'C' is allowed only the fourth position with low frequency. MYOGNF1_01 is a matrix associated with myogenin, nuclear factor 1 or related factors, and is therefore involved in the regulation of differentiation. CREL_01 is a matrix associated with the C-Rel proto-oncogene protein (C-Rel protein). An understanding of the function of these factors is important to this study. The PWM groups described above may be involved in tissue-specific gene regulation. If all housekeeping genes have CGI [25] then genes without CGI can be assumed to be tissue-specific or rarely expressed. Thus, genes with a cluster of predicted TFBS not associated with CGI might be associated with tissue-specific regulation. Further analysis of extractions of tissue specific genes, shown in Results, supports the hypothesis.

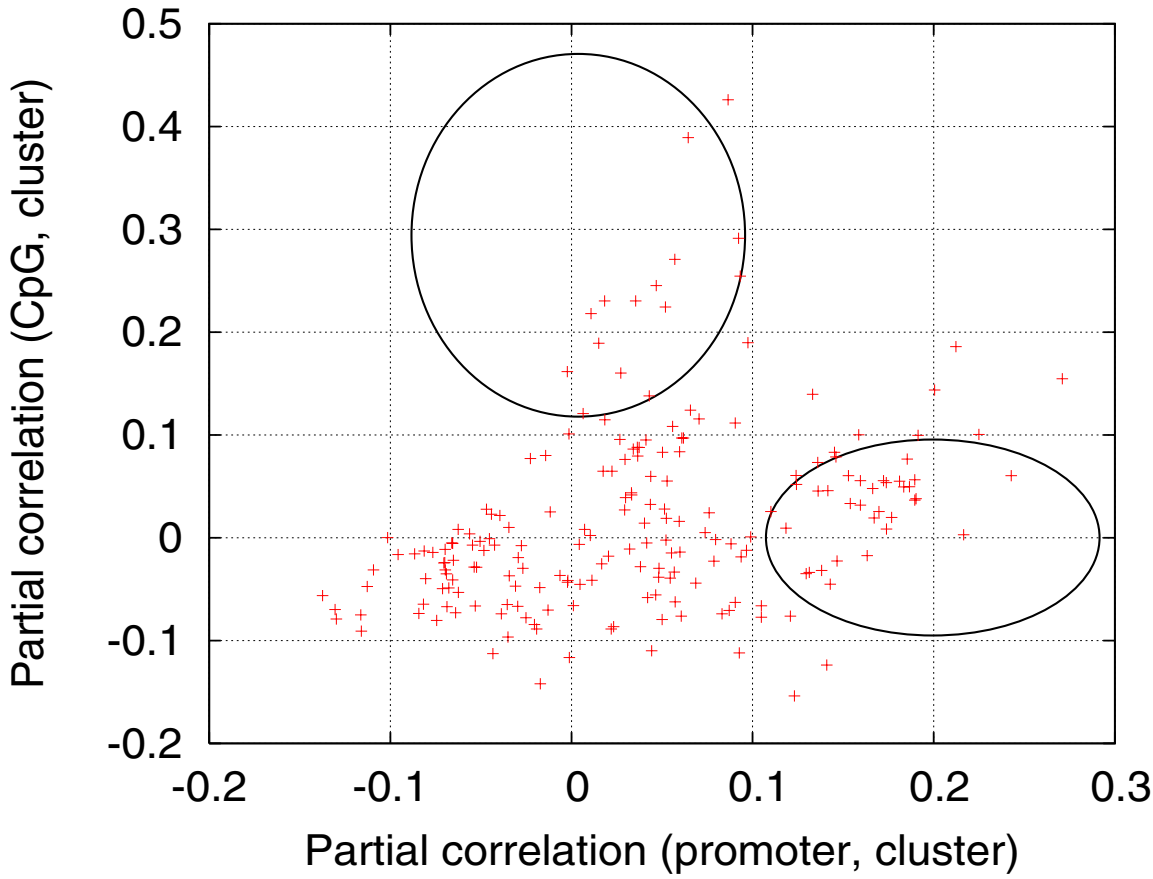

Figure 4
Title: Plot of $r_{IC.P}$ against $r_{PC.I}$ for various PWMs. The top circle is the area where $r_{PC.I}$ is around zero and $r_{IC.P}$ is high. The right circle is the area where $r_{IC.P}$ is around zero and $r_{PC.I}$ is high. The two circles were drawn manually. Ideal CGI-related and CGI-independent PWMs are to be plotted in the top and right circles, respectively.

Results from this analysis provide a solution to the promoter prediction problem. Hannenhalli *et al.* used additional information, including profiles of TF binding sites, for promoter prediction based on CGI [29], with no significant improvement to prediction performance. The report using 7 manually selected PWMs confirmed that CGI are the most dominant feature. Our results show that Sp1 and ATF have a strong correlation with CGI; a result consistent with their result that information including both PWM did not improve prediction accuracy. This observation is consistent with other PWMs. More stringent selection of PWMs is required for an improved accuracy of promoter prediction. One strategy is to utilise the CGI-independent PWMs identified in this study. Another problem is exemplified by the under-representation of Oct-1 (M00138) in the (-600:600) region of the human promoter and the absence of positional preferences [29]. This under-representation was not expected but is observed in 10% of known PWMs. OCT1_04 (M00138) is not in the high quality list of TRANSFAC, OCT1_01 (M00135) and OCT1_C (M00210) was found to have minus cluster scores (-0.63 and -2.89) in our table (additional file 1).

It is noteworthy that Fig. 5 shows TFBS (AP2_Q6) in non-promoters to occur randomly under a certain distribution. This distribution can be modelled by a binomial probability distribution. A model of Poisson distribution, which is an approximation of binomial probability distribution for a certain condition, was proposed in [11] as the probability distribution of TFBS density. Although we have not tested the goodness-of-fit, our observation does not contradict the Poisson distribution model.

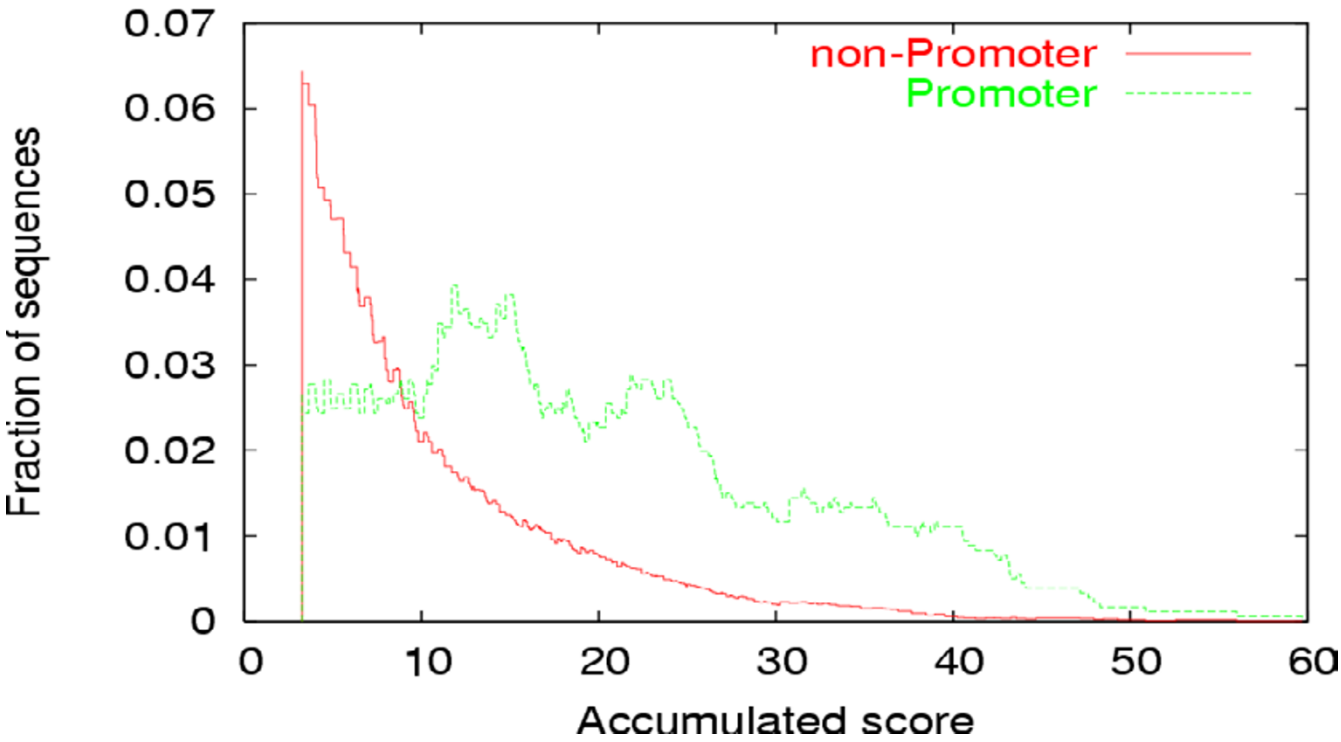

Figure 5
Title: Distribution of accumulated score C for promoters and non-promoters for AP2_Q6

To assess the robustness of the cluster score, we compared cluster scores for different datasets from chromosomes 20, 21 and 22 (Fig. 3). The correlation coefficients were 0.91 (a) and 0.93 (b), proving that the significance would be similar if we utilized the whole human genome dataset in the analysis. The scale of the figures between the Y-axis and X-axis are different because of the different number of sequences taken from each chromosome.

Conclusions

We have developed a measure that statistically evaluates the degree of concentration of predicted TFBS in promoter sequences. Using this strategy to analyse various PWMs we have determined that predicted TFBS tend to cluster in human promoter sequences rather than in non-promoter sequences. Our results show that local concentrations of predicted TFBS in human promoter sequences are not a general characteristic of PWMs. Only a portion of identified PWM matches corresponded to TFBS occurring in clusters in promoter sequences. By computing partial correlation coefficients, we identified PWM sets associated with CGI and others that are independent of CGI. Transcription factors and binding sites associated with CGI-independent PWMs are likely to be involved in tissue-specific gene regulation. Indeed, using the CGI-related/dependent PWM sets, we extracted tissue-specific genes with high accuracy by detecting clusters of predicted TFBS. These results will be useful to interpret predicted transcription factor binding sites and to further understand the role of their formation into clusters. Ultimately, these findings will further elucidate the various functions of promoters, genes and transcription factors.

Methods

Data

DNA sequences from the fully sequenced chromosomes (chromosomes 20, 21 and 22) were taken from the November, 2002 GenBank freeze (build 31) and assembled by NCBI, in accordance with the annotation of the UCSC genome browser [30]. RefSeq [31] genes were used as they have been reviewed by NCBI staff, are well studied, and are unlikely to be spurious. Some genes in the human genome have alternative promoters [32], complicating our analysis. For this reason, overlapping genes identified using the UCSC annotation were discarded. This check of RefSeq genes reduced the number of genes in the analysis from 527 to 373 for chromosome 20, 224 to 142 for chromosome 21, and 449 to 294 for chromosome 22. The resultant gene set *U* consists of 809 genes.

Table 2: CGI-related PWMs in descending order of $r_{IC.P}$ = *Y*. The columns are: rank number, accession number, Identifier in TRANSFAC, $r_{PC.I}$ (=*X*), $r_{IC.P}$ (=*Y*), cluster score and threshold.

Rank	ACCESSION	ID	X	Y	S	T
1	M00332	WHN_B	0.09	0.43	158.4	1.9
2	M00736	E2F1DP1_01	0.06	0.39	151.8	2.6
3	M00739	E2F4DP2_01	0.09	0.29	91.4	0.9
4	M00737	E2F1DP2_01	0.06	0.27	81.9	0.9
5	M00108	NRF2_01	0.09	0.25	72.6	0.9
6	M00055	NMYC_01	0.05	0.25	34.6	0.9
7	M00235	AHRARNT_01	0.02	0.23	26.8	0.9
8	M00740	E2F1DP1RB_01	0.04	0.23	48.1	0.9
9	M00652	NRF1_Q6	0.05	0.22	105.3	0.9
10	M00466	HIF1_Q5	0.01	0.22	19.7	0.9
11	M00341	GABP_B	0.1	0.19	46.6	0.9
12	M00738	E2F4DP1_01	0.02	0.19	28.6	0.9
13	M00538	HTF_01	0	0.16	9.7	0.8
14	M00694	E4F1_Q6	0.03	0.16	23.6	0.9
15	M00743	CETS168_Q6	0.13	0.14	47.1	1
16	M00650	MTF1_Q4	0.04	0.14	22.6	0.9
17	M00243	EGR1_01	0.07	0.12	32.4	0.9
18	M00251	XBP1_01	0.01	0.12	7.8	0.9
19	M00691	ATF1_Q6	0.07	0.12	17.3	0.9
20	M00236	ARNT_01	0.02	0.11	6.5	1
21	M00143	PAX5_01	0.09	0.11	25.7	0.8
22	M00273	R_01	0.06	0.11	23.8	0.8
23	M00244	NGFIC_01	0.06	0.1	23	0.9
24	M00280	RFX1_01	0.06	0.1	11.1	0.9
25	M00121	USF_01	0.03	0.1	7.6	1
26	M00287	NFY_01	0.04	0.1	21.3	1.9
27	M00039	CREB_01	0.04	0.09	23.2	1
28	M00309	ACAAT_B	0.04	0.09	6.8	0.9
29	M00651	NFMUE1_Q6	0.03	0.09	13	1.8
30	M00017	ATF_01	0.06	0.08	19.2	1
31	M00481	AR_01	0.05	0.08	7.5	0.8
32	M00041	CREBP1CJUN_01	0.04	0.08	20.4	1
33	M00040	CREBP1_01	0.03	0.08	4.7	0.9
34	M00114	TAXCREB_01	0.02	0.06	7.3	0.9
35	M00279	MIF1_01	0.02	0.06	10.9	1.8
36	M00246	EGR2_01	0.04	0.06	9.7	0.9
37	M00085	ZID_01	0.05	0.06	8	0.8

To increase the accuracy of the annotation of transcriptional start sites, we modified the annotation of RefSeq according to DBTSS (version 2, Mar 2002) [33], a database of transcriptional start sites for 5' end mRNA sequences. Suzuki *et al.* reported that a certain portion of sequences in DBTSS were longer (extended) toward 5' end of mRNA sequences than those in RefSeq [33]. We describe how the modification improved the first gene set *U*. Fig. 6 shows the composition of different gene collections. The RefSeq database is updated daily by increasing the number of entries and correcting others. For illustration purpose, two versions of RefSeq are shown in Fig. 6. The old RefSeq is the version analysed by Suzuki *et al.* The new RefSeq is the current version used in this study. Of the 217,402 sequences contained in DBTSS 7,889 correspond to sequences in RefSeq and are referred to as cloned RefSeq sequences. The extension rate, defined as the rate of extension of mRNAs sequences from cloned RefSeq by DBTSS, was 0.34. Therefore, $|\{D,G\}|$ (the number of gene set $\{D,G\}$) = 7,889 genes, $|\{Dex,Gex\}|$ = 2,683 genes and $|\{Dex,Gex\}|/|\{D,G\}|$ = 0.34. The ftp site of DBTSS provides the set of extended mRNA sequences (I_{ftp} = $\{Dex,Gex\}$). The gene set from chromosomes 20, 21 and 22 is a partial set $U = \{C^u, E^u, F^u, G_n^u, G_{ex}^u\}$ from New RefSeq. We can identify the genes in set G_{ex}^u as the conjunction of I_{ftp} and *U*. The number of G_{ex}^u was counted (273

Table 3: CGI-independent PWMs in descending order of $r_{PC.I}$ (=*X*). The columns are: rank number, accession number, Identifier in TRANSFAC, $r_{PC.I}$ (=*X*), $r_{IC.P}$ (=*Y*), cluster score and threshold.

Rank	ACCESSION	ID	X	Y	S	T
1	M00491	MAZR_01	0.27	0.15	117.4	1.8
2	M00706	TFIII_Q6	0.24	0.06	52.7	3.5
3	M00324	MINI20_B	0.22	0.1	53.2	0.8
4	M00056	MYOGNF1_01	0.22	0	31.6	1.3
5	M00649	MAZ_Q6	0.21	0.19	114.4	3.7
6	M00665	SP3_Q3	0.2	0.14	67.7	1.7
7	M00032	CETS1P54_01	0.19	0.1	47.7	1.8
8	M00053	CREL_01	0.19	0.04	26.9	0.8
9	M00054	NFKAPPAB_01	0.19	0.06	33.5	0.9
10	M00632	GATA4_Q3	0.19	0.04	25.1	0.6
11	M00373	PAX4_01	0.19	0.05	26.1	0.6
12	M00072	CP2_01	0.19	0.08	32	0.9
13	M00733	SMAD4_Q6	0.18	0.05	26.3	0.8
14	M00134	HNF4_01	0.18	0.06	25.7	0.6
15	M00194	NFKB_Q6	0.18	0.02	28.5	0.8
16	M00445	XVENT1_01	0.17	0.01	19.9	0.7
17	M00057	COMP1_01	0.17	0.05	24.1	0.5
18	M00097	PAX6_01	0.17	0.06	24.1	0.5
19	M00104	CDPCR1_01	0.17	0.03	21.3	0.6
20	M00222	HAND1E47_01	0.17	0.02	20.4	0.8
21	M00626	EFC_Q6	0.17	0.05	22.6	0.6
22	M00745	LEF1_Q6	0.16	-0.02	15.9	0.8
23	M00707	TFIIA_Q6	0.16	0.03	20.2	0.7
24	M00086	IK1_01	0.16	0.06	24.1	0.9
25	M00329	PAX9_B	0.16	0.1	33.7	0.7
26	M00478	CDC5_01	0.15	0.03	19	0.6
27	M00670	TCF1P_Q6	0.15	0.06	22.7	0.8
28	M00257	RREB1_01	0.15	-0.02	15.8	0.8
29	M00007	ELK1_01	0.15	0.08	31	0.8
30	M00698	HEB_Q6	0.15	0.08	28.7	0.9
31	M00052	NFKAPPAB65_01	0.14	-0.05	9.4	0.9
32	M00514	ATF4_Q2	0.14	0.05	21.8	1.7
33	M00191	ER_Q6	0.14	-0.03	11	0.8
34	M00003	VMYB_01	0.14	0.05	18	0.8
35	M00261	OLF1_01	0.14	0.07	24.6	0.8
36	M00490	BACH2_01	0.13	-0.03	9.3	0.7
37	M00001	MYOD_01	0.13	-0.03	10.4	0.9
38	M00634	GCM_Q2	0.12	0.05	19.8	0.8
39	M00035	VMAF_01	0.12	0.06	17.5	0.7
40	M00340	ETS2_B	0.12	-0.08	5	0.8
41	M00005	AP4_01	0.12	0.01	14.1	0.8
42	M00701	SMAD3_Q6	0.11	0.03	11.4	0.8
43	M00531	NERF_Q2	0.1	-0.08	4.8	0.9
44	M00339	ETS1_B	0.1	-0.07	5.7	0.9
45	M00657	PTF1BETA_Q6	0.1	0	7.5	0.9
46	M00254	CAAT_01	0.1	-0.01	6.6	0.9
47	M00118	MYCMAX_01	0.09	-0.02	6.2	0.9
48	M00693	E12_Q6	0.09	-0.01	6.5	0.9
49	M00004	CMYB_01	0.08	0	7.1	0.9
50	M00238	BARBIE_01	0.08	0.02	9.4	0.9
51	M00648	MAF_Q6	0.07	0.01	5.8	0.8
52	M00002	E47_01	0.06	0.02	5.3	0.9
53	M00262	STAF_01	0.05	0	9.2	0.9
54	M00119	MAX_01	0.05	0.03	4.9	1

Table 4: The gene list sorted by *DCC* score. The genes, in which clusters of TFBS are found on promoters using CpG-related/ independent PWMs, and tissue specificity, have been previously identified. HK denotes housekeeping. Tissue specific genes can be selected independent of CpG islands (start_p) using *DCC* score.

1	NM006272	0.43	brain	0
2	NM007341	0.4	muscle	0
3	NM002592	0.37	brain	0.86
4	NM001819	0.27	brain	0.68
5	NM004414	0.23	kidney	0.89
6	NM002999	0.19	kidney	0.73
7	NM003195	0.16	brain	0.73
8	NM002591	0.14	liver	0
9	NM000454	0.11	HK.liver	0.87
10	NM003312	0.1	liver	0.72
11	NM004339	0.09	brain	0.9
12	NM020708	0.08	brain	0.64
13	NM006870	0.05	HK	0.7
14	NM003277	0.04	lung	0.74
15	NM005194	0.04	brain	0.86
16	NM003610	0.01	brain	0.76
17	NM000355	-0.03	kidney	0
18	NM002430	-0.03	muscle	0.75
19	NM006767	-0.03	brain	0.74
20	NM005137	-0.03	muscle	0.76
21	NM003279	-0.05	muscle	0
22	NM004535	-0.05	brain	0
23	NM007019	-0.05	HK	0.72
24	NM013236	-0.07	HK	0.69
25	NM004175	-0.07	brain	0.72
26	NM001958	-0.07	muscle	0
27	NM001338	-0.13	vulva	0.84
28	NM002676	-0.14	HK	0.63
29	NM003098	-0.16	muscle	0.71
30	NM002854	-0.17	brain	0
31	NM002305	-0.23	HK	0
32	NM005080	-0.25	HK	0.84
33	NM001024	-0.25	HK	0.76
34	NM021974	-0.26	HK	0.63
35	NM014876	-0.3	HK	0.95
36	NM001098	-0.34	muscle	0.65
37	NM000071	-0.37	liver	0.8
38	NM006198	-0.37	brain	0
39	NM001675	-0.39	HK.muscle	0.8
40	NM005423	-0.68	brain	0

genes). If the extension rate of mRNAs sequences for $\{G_n^u, G_{ex}^u\}$ is also 0.34, then $|\{G_{ex}^u\}|/|\{G_n^u, G_{ex}^u\}| = 0.34$. Therefore, $\{|G_n^u, G_{ex}^u\}|$ is estimated to be 802.9 genes. As $|U| = 809$, the number of genes in $V = \{C^u, E^u, F^u\}$ is estimated to be 6.1 genes. If all human genes are cloned by the cap-targeted selection method called oligo-capping, a greater number will have extended 5' ends. If this were the case then 34% (or 2.1 genes) of *V* would have extended 5' ends. Thus, any further correction of TSS for our gene set is expected to be quite small.

We identified the conjunction set G_{ex}^u (273 genes) of the above collected 809 RefSeq genes and 2,683 genes in DBTSS. The set G_{ex}^u was examined to determine if extended sequences existed on human genome sequences and if they are registered in the new RefSeq. Of this set, the BLAT program[34] identified 30 genes in which the 5' end sequences could not be detected. Due to the uncertainty of TSS these genes were not used in this study. Forty-one DBTSS mRNA sequences from G_{ex}^u were shorter than corresponding sequences in the new RefSeq with regard to 5' end sequences. It is assumed that these RefSeq sequences

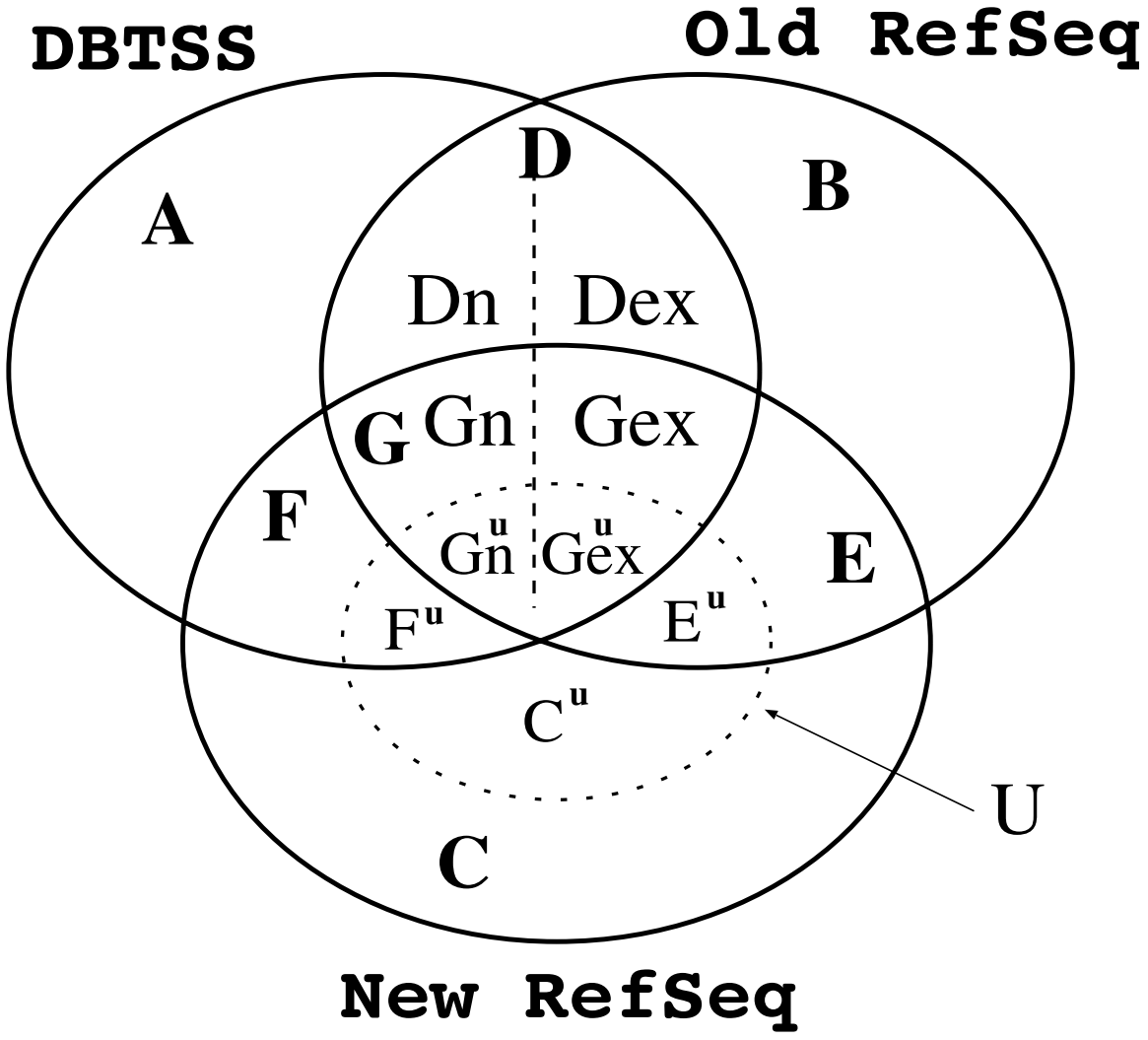

Figure 6

A Venn diagram of three gene sets (DBTSS, old RefSeq, and new RefSeq). Gene sets from *A* to *G* (Bold alphabet) consist of genes in the regions bounded by the thick lines. *D* consists of *Dn* (genes whose 5' end sequences were not extended from the old RefSeq sequences with DBTSS data) and *Dex* (genes whose 5' end sequences were extended). *G* consists of *Gn* (genes whose 5' end sequences were not extended from the old RefSeq sequences with DBTSS data) and *Gex* (genes whose 5' end sequences were extended). Namely *D* = {*Dn*,*Dex*} and *G* = {*Gn*,*Gex*}. Genes in chromosomes 20, 21, 22 were denoted by *U* = {C^u, E^u, F^u, G_n^u , G_{ex}^u }. Gene sets C^u, E^u,F^u, G_n^u and G_{ex}^u are parts of *C*, *E*, *F*, *Gn* and *Gex*, respectively. Some of the numbers of the sets are given in [33], that is, |{*D*,*G*}| = 7889, |{*Dex*,*Gex*}| = 2683 and |{*Dex*,*Gex*}| / |{*D*,*G*}| = 0.34, where |{*D*,*G*}| denotes the number of genes in set {*D*,*G*}.

were corrected following the old RefSeq release independent of DBTSS and were used as they were. Finally, we modified the exon annotation of 202 genes according to DBTSS.

We extracted promoter sequences at relative positions (-600:-1) from the TSS, and intron subsequences 600 bp in length from genome sequences. Only intron sequences were used for the non-promoter sequence data sets as exon sequences are known to have preferences in their oligomer statistics, such as G+C % and codon bias [35]. The first intron was not included in the data set as although regulatory elements are rare in introns, intron 1 occasionally contains regulatory elements such as enhancers. We

investigated the frequency of enhancers in human introns by searching NCBI PubMed [36] with the keywords 'human', 'first intron', and 'enhancer'. This search yielded 194 papers. Replacing the keyword 'first intron' with 'second', 'third', 'fourth intron', 'fifth' or 'last intron' yielded 40, 15, 1, 1 and 6 papers, respectively. Replacing 'enhancer' with 'silencer' resulted in 281, 6, 3, 0, 0 and 0 papers, respectively. Removal of intron 1 from the data set greatly reduces the overall occurrence of regulatory elements in human intron sequences and allows our statistical analysis to be performed without significant interference from intronic regulatory sequences. Intergenic sequences are left out of the non-promoter dataset due the unknown occurrence of regulatory sequences.

Prediction of TFBS

Each promoter or non-promoter sequence was scanned by the MATCH program using 423 matrices in TRNASFAC version 6.3 (a transcription factor database) with options including 'vertebrate', 'minimize false negatives' (in cut-off selection) and 'use high quality matrices only'. As Kel *et al.* described, the cut-off was determined so that the false negative rate is 10% [3]. The option 'use high quality matrices only' uses approximately 70% of matrice [3]. Any PWM in the 'high quality' PWMs meet the criteria; When the PWM is used with a cut-off value which allows a false negative rate of 50%, then the match rate dropped below 1 match/kb in exon2 sequences [3]. If more than one matrix was matched to same transcription factor (prefix of "Identifier"), we selected a representative matrix with the highest quality and smallest suffix number according to the TRANSFAC definition. After scanning the sequences by MATCH, we set consecutive sampling windows (600 bp) in introns and promoter sequences, and then recorded corresponding TFBS predictions. To prevent double counting of palindromic binding sites, two matches for the same matrix at the same position was regarded as a single match and the match with the higher score was taken. Before MATCH ran, repeat sequences were masked to 'N' according to the annotation by RepeatMasker in the UCSC genome browser. From the above analysis we extracted 361, 129, and 278 promoter sequences from chromosomes 20, 21 and 22, respectively. The promoter sequences identified contained repeat sequences (e.g. ALU, L1) and simple repeats with low complexity, as observed in intron sequences. These sequences account for about 20% of all bases. To balance the rate of repeats between promoters and introns, we discarded intron sequences with high rates of repeats, so that the average rate of repeats in the intron samples was at the same level as in promoter sequences. The number of 600 bp intron sequences included in the analysis was 6,589 (chromosome 20), 4,324 (chromosome 21) and 4,531 (chromosome 22).

Accumulated scores of TFBS

When predicted TFBS occur many times in a sequence there is a high probability that it contains functional regulatory regions or promoters [7,8,19,28]. We tested this hypothesis for individual PWMs. The degree of concentration of predicted TFBS in a sequence was defined as the accumulated score *C*, which is a summation of the MATCH score for PWMs in the subsequence and is calculated for each PMW and corresponding sequence. *C* is assumed to be almost proportional to the frequency of predicted TFBS for the corresponding PWM. Many sequences generate different *C* values although some are identical. We then generated a series of C_j (j = 1 ... n) values for a PWM, where n is the number of different *C* values. Fig. 5 shows the histogram of *C* for promoters and non-promoters respectively, using the TRANSFAC matrix of identifier 'AP2_Q6' as an example. Since *C* reflects the number of predicted TFBS found in a sequence, the figure shows the density of predicted TFBS in a sequence. This result is similar to the density plot described by Pestridge and Burks [7], although our figure (Fig. 5) is not a plot of predicted TFBS density for mixed PWMs, but instead is a plot of predicted TFBS density for individual PWMs. Also, the X-axis in our plot does not indicate the number of predicted TFBS but instead indicates the accumulated score *C*. The Y-axis is smoothed by averaging for the width of 5 in *C* value.

Cluster score and statistical significance for a PWM

Significance values for an individual PWM from a series of C_j can be determined from a contingency table. Table 5 shows a contingency table for the number of promoters and non-promoters above and below the threshold C_j for a given PWM. From this table, χ^2 value for a given C_j is defined as

Table 5: A contingency table when predicted TFBS and a threshold *T* are given.

	Sequences where TFBS clusters found	Sequences where TFBS clusters not found	Sum
# of promoter	A_1	A_2	A
# of non-promoter	B_1	B_2	B

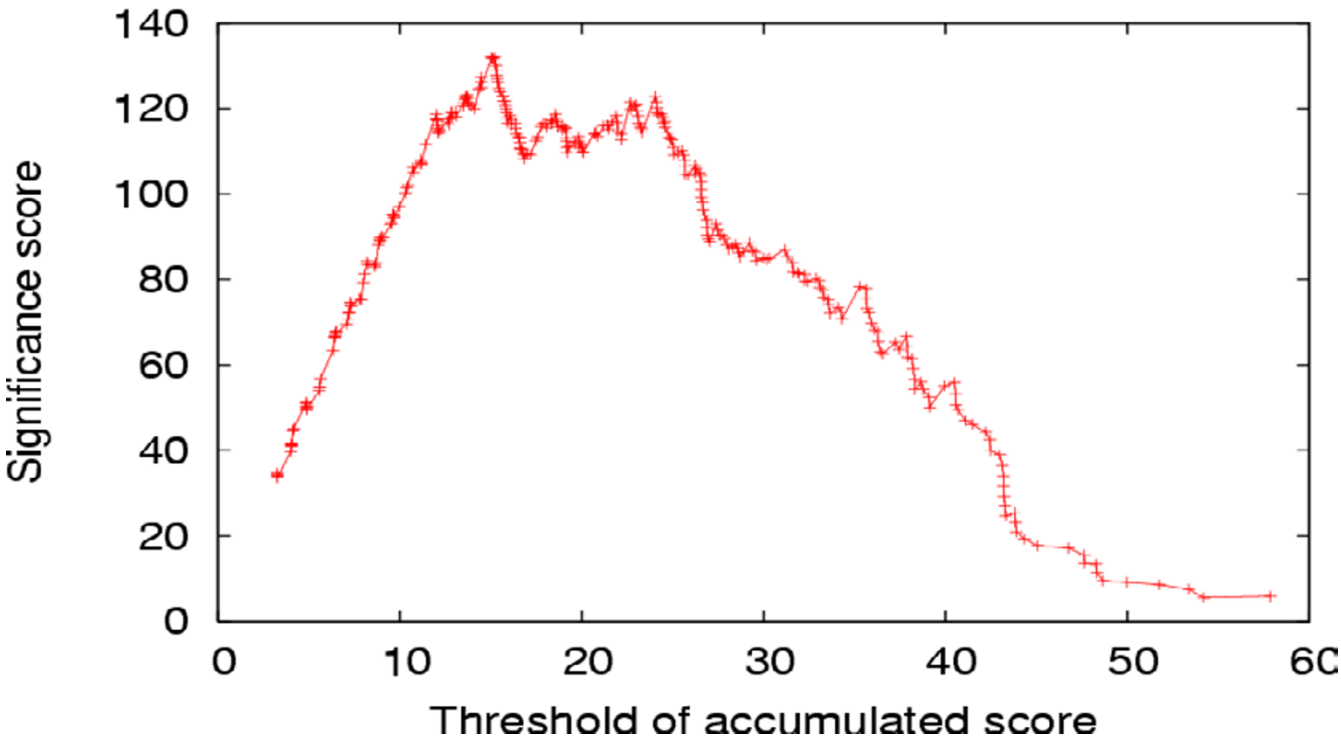

Figure 7
Title: significant score Q_j of matrix AP2_Q6 for different thresholds.

$$\chi_j^2 = \sum_{i=1}^{2} \frac{(B_i\sqrt{A/B} - A_i\sqrt{B/A})^2}{A_i + B_i}, \qquad (2)$$

where

$$A = \sum_{i=1}^{2} A_i, B = \sum_{i=1}^{2} B_i$$

described in [37]. From the χ^2 value, we computed the probability P that the χ^2 value or greater is obtained by chance. The probability P was calculated from the χ^2. Since P is calculated for many PWMs, we must deal with the problem of multiple testing. Using the Bonferroni correction [23], Pn was calculated using the formula $1 - P_n = (1 - P)^n$, approximately $P_n = P \times n$ for small $P \times n$. The n is the number of PWMs. When we determine the set of significant PWMs, Pn were compared with the significance level (i.e. 0.01). We also defined the statistical significance Q_j as $Q_j = -\log_{10}(P_n)$ if $Rprom > Rnonprom$ and $Q_j = +\log_{10}(P_n)$ otherwise, where $Rprom = A1/A$ (a rate of sequences in promoters where clusters found), $Rnonprom = B1/B$ (a rate of sequences in non-promoters where clusters found). Although the P is an indicator of the difference between the occurrence of promoters and non-promoters, the probability P itself does not represent the preferences of PWMs for promoters. To represent the preference of predicted TFBS for or against promoters, we add signs for statistical significance Q_j. Positive Q_j indicates that predicted TFBS tend to appear frequently in promoters, while negative Q_j indicates that predicted TFBS tend to avoid promoters.

We studied how statistical significance Q_j varies with the threshold of $C\,j$. Fig. 7 shows the presence of a peak of Q_j when we change the threshold. We define the cluster score S of a PWM in such a way that the significance is the maximum, namely

$$S = \max_j \{|Q_j|\}. \qquad (3)$$

We simultaneously define a unique threshold T of the PWM by

$$T = \operatorname*{argmax}_{C_j} |Q_j|. \qquad (4)$$

For the all-vertebrate TRANSFAC PWMs, we determined thresholds T and calculated significance scores (or cluster scores) S. The highest scoring PWMs are listed in Table 1.

Correlations among promoter sequences, CpG islands, and clusters

For every 600 bp sequence, three numerical features (promoter, CGI, and clusters) were annotated. CGI were identified using the CpGProD program [38] for original long sequences (not for short sequences of 600 bp). Regions larger than 500 bp with a G+C % equal to or greater than 50% and 'observed CpG / expected CpG' equal to or greater than 0.60 were classified as CGI [38,39]. The CpGProD program outputs 'start_p' scores for the predicted CGI. This score indicates the probability that the region is a CGI located over a transcription start site (start CGI). Short 600 bp sequences sampled from long sequences containing CGI were annotated as CGI if the overlapping CGI region was longer than 300 bp. The accumulated score C was used for cluster annotation. From sequences with feature annotation, the correlation coefficients between every two of the three features were computed for each PWM by the statistical language R [40]. We use P to denote whether the sequences is promoter or not, namely $P = \{1,0\}$, and I to the denote 'start_p' score for CGI calculated using the CpGProD program [38]. A partial correlation coefficient for each PWM was calculated using the subsequences. For example, a partial correlation coefficient $r_{PC.I}$ is the correlation between P and C while controlling for I, defined by

$$r_{PC.I} = \frac{r_{PC} - r_{PI} r_{CI}}{\sqrt{1 - r_{PI}^2}\sqrt{1 - r_{CI}^2}}, \qquad (5)$$

where r_{PI} is a correlation coefficient between variable P and I, r_{PC} is a correlation coefficient between variable P and C and r_{CI} is a correlation coefficient between variable C and I. A partial correlation coefficient differs from a correlation coefficient. If the correlation between P and C depends entirely on the common cause I, then when I is constant, the correlation between P and C should be zero. The partial correlation $r_{PC.I}$ expresses such a relationship. Even when I varies, $r_{PC.I}$ is expected to be zero in such a situation, while the correlation coefficient r_{PC} may not be zero. See chapter 16.4 in [41] for details.

Fig. 4 shows a plot of $r_{IC.P}$ against $r_{PC.I}$ for various PWMs. Using these two values, we identified two PWM sets including, (1) a CGI-related set consisting of 37 PWMs in which the clusters are correlated with CGI independent of promoters, and (2) a CGI-independent set consisting of 54 PWMs, in which the clusters are correlated with promoters independent of CGI. The CGI-related set requires that $r_{IC.P} > r_{PC.I}$ and that the partial correlation coefficient (PCC) $r_{IC.P}$ is significantly high ($p < 0.01$) under the hypothesis that $r_{IC.P}$ is zero (see below). The CGI-independent set requires that $r_{IC.P} < r_{PC.I}$ and that $r_{PC.I}$ is likewise significantly high. To calculate the statistical significance of PCC, a PCC r was subjected to z-transformation defined as

$$z = \ln\left(\frac{1+r}{1-r}\right). \qquad (6)$$

The values of z are supposed to be normally distributed and the expected variance is

$$\sigma_z = \frac{1}{n-3}, \qquad (7)$$

where n is the sample size [41]. The CGI-related PWMs and CGI-independent PWMs are listed in Tables 2 and 3.

Gene expression data

To examine the relationship between clusters of predicted TFBS and the tissue specificity of the genes where clusters were found, we generated a list of genes with expression data. This list includes 535 housekeeping or maintenance genes expressed in 11 human adult and foetal tissues from [26], and 451 housekeeping or maintenance genes available at HugeIndex database http://www.hugeindex.org[27]. We then identified 581 non-redundant housekeeping genes and 'tissue-selective' genes, which are predominantly, but not exclusively, expressed in one tissue type. Tissue-selective genes were expressed in brain (589 genes), kidney (129 genes), liver (271 genes), lung (68 genes), muscle (302 genes), prostate (45 genes) and vulva (95 genes) [27]. These genes corresponded to 2,069 RefSeq entries. Seventy-two of these genes were identified in our gene set covering chromosome 20, 21 and 22 and were used for further analysis.

Tissue specific gene detection based on clusters of predicted TFBS

Using the two PWM sets described above, we searched clusters of predicted TFBS in promoter sequences. We calculated C and statistical significance of C as follows. For each promoter sequence, 30 random 600 bp sequences were generated under the first order Markov model, which is based on dinucleotide frequency and can identify promoter CpG bias and G+C%. The MATCH program was run with given matrix thresholds. The accumulated score C of PWMs was computed in every random sequence. A mean $\overline{C}_M$ and a variance σ_M^2 of C for each PWM were estimated from the random sequences. Then, for the given promoter sequence S_p and the PWM (M), we can run MATCH and calculate C and its significance score (Z-score), namely

$$Z(S_p, M) = \frac{(C - \bar{C}_M)}{\sigma_M}. \qquad (8)$$

If the *Z*-score value was above three, the promoter sequence was taken and considered together with the PWM. Table 4 lists genes identified with clusters of predicted TFBS with significant *C* values for the CGI-related/dependent PWM set. The computation of p-value of cumulative hypergeometric distribution was performed using the AS R77 algorithm [42].

Authors' contributions

KM designed the study and carried out statistical analysis. TK participated in the design and carried out functional analysis. YS directed the study. All authors read and approved the final manuscript.

Additional material

Additional File 1

The list of PWM-PCP/NCP sorted by cluster score. Each column represents rank number, accession number in TRANSFAC, identifier in TRANSFAC, cluster score, threshold.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-16-S1.csv]

References

1. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
2. 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.** *Nucleic Acids Res* 1995, **23:**4878-4884.
3. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3576-3579.
4. Lenhard B, Wasserman WW: **TFBS: Computational framework for transcription factor binding site analysis.** *Bioinformatics* 2002, **18:**1135-1136.
5. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:**374-378.
6. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26:**225-228.
7. Prestridge DS, Burks C: **The density of transcriptional elements in promoter and non-promoter sequences.** *Hum Mol Genet* 1993, **2:**1449-1453.
8. Arnone MI, Davidson EH: **The hardwiring of development: organization and function of genomic regulatory systems.** *Development* 1997, **124:**1851-1864.
9. Prestridge DS: **Predicting Pol II promoter sequences using transcription factor binding sites.** *J Mol Biol* 1995, **249:**923-932.
10. Solovyev V, Salamov A: **The Gene-Finder computer tools for analysis of human and model organisms genome sequences.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**294-302.
11. Wagner A: **Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.** *Bioinformatics* 1999, **15:**776-784.
12. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci U S A* 2002, **99:**757-762.
13. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:**1019-1028.
14. Papatsenko DA, Makeev VJ, Lifanov AP, Regnier M, Nazina AG, Desplan C: **Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers.** *Genome Res* 2002, **12:**470-481.
15. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11:**1559-1566.
16. Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17:**878-889.
17. Frith MC, Spouge JL, Hansen U, Weng Z: **Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences.** *Nucleic Acids Res* 2002, **30:**3214-3224.
18. Levy S, Hannenhalli S: **Identification of transcription factor binding sites in the human genome sequence.** *Mamm Genome* 2002, **13:**510-514.
19. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for comparative sequence-based discovery of functional transcription factor binding sites.** *Genome Res* 2002, **12:**832-839.
20. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: **Toucan: deciphering the cis-regulatory logic of coregulated genes.** *Nucleic Acids Res* 2003, **31:**1753-1764.
21. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo.** *BMC Bioinformatics* 2002, **3:**30.
22. Johansson O, Alkema W, Wasserman WW, Lagergren J: **Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm.** *Bioinformatics* 2003, **19 Suppl 1:**I169-I176.
23. Ewens WJ, Grant GR: **Statistical Methods in Bioinformatics: An Introduction.** *Springer-Verlag New York, Inc.*; 2001.
24. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences.** *Nucleic Acids Res* 1990, **18:**6097-6100.
25. Larsen F, Gundersen G, Lopez R, Prydz H: **CpG islands as gene markers in the human genome.** *Genomics* 1992, **13:**1095-1107.
26. Warrington JA, Nair A, Mahadevappa M, Tsyganskaya M: **Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes.** *Physiol Genomics* 2000, **2:**143-147.
27. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, Weng Z, Mutter GL, Frosch MP, Macdonald ME, Milford EL, Crum CP, Bueno R, Pratt RE, Mahadevappa M, Warrington JA, Stephanopoulos G, Gullans SR: **A compendium of gene expression in normal human tissues.** *Physiol Genomics* 2001, **7:**97-104.
28. Levy S, Hannenhalli S, Workman C: **Enrichment of regulatory signals in conserved non-coding genomic sequence.** *Bioinformatics* 2001, **17:**871-877.
29. Hannenhalli S, Levy S: **Predicting transcription factor synergism.** *Nucleic Acids Res* 2002, **30:**4278-4284.
30. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
31. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence project: update and current status.** *Nucleic Acids Res* 2003, **31:**34-37.
32. Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, Okubo K, Sakaki Y, Nakamura Y, Suyama A, Sugano S: **Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites.** *EMBO Rep* 2001, **2:**388-393.

33. Suzuki Y, Yamashita R, Nakai K, Sugano S: **DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs.** *Nucleic Acids Res* 2002, **30:**328-331.
34. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
35. Claverie JM, Sauvaget I, Bougueleret L: **K-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping.** *Methods Enzymol* 1990, **183:**237-252.
36. **PubMed.** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed]
37. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: **Numerical recipes in C.** secondth edition. *New York, Press Syndicate of the University of Cambridge*; 1992.
38. Ponger L, Mouchiroud D: **CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences.** *Bioinformatics* 2002, **18:**631-633.
39. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes.** *J Mol Biol* 1987, **196:**261-282.
40. Ihaka R, Gentleman R: **A Language for Data Analysis and Graphics.** *Journal of Computational and Graphical Statistics* 1996, **5:**299-314.
41. Sokal RR, Rohlf FJ: **Biometry.** Thirdth edition. *Freeman and Company*; 1995.
42. Shea B, Remark AS: **AS R77: A remark on algorithm AS 152: Cumulative hypergeometric probabilities.** *Applied Statistics* 1989, **38:**199-204.