DNA Research, 2018, 25(1), 103–112
doi: 10.1093/dnares/dsx041
Advance Access Publication Date: 23 October 2017
Full Paper

OXFORD

Full Paper

Investigating transcription factor synergism in humans

Fabio Cumbo[1,2,3], Davide Vergni[4], and Daniele Santoni[1,*]

[1]Institute for Systems Analysis and Computer Science 'Antonio Ruberti', National Research Council of Italy, 00185 Rome, Italy, [2]Department of Engineering, Third University of Rome, 00146 Rome, Italy, [3]SYSBIO.IT–Centre of Systems Biology, 20126 Milan, Italy, and [4]Institute for Applied Mathematics 'Mauro Picone', National Research Council of Italy, 00185 Rome, Italy

*To whom correspondence should be addressed. Tel. +39 0649937119. Fax. +39 0649937137.
Email: daniele.santoni@iasi.cnr.it

Edited by Dr. Mikita Suyama

Received 22 February 2017; Editorial decision 15 September 2017; Accepted 19 September 2017

Abstract

Proteins are the core and the engine of every process in cells thus the study of mechanisms that drive the regulation of protein expression, is essential. Transcription factors play a central role in this extremely complex task and they synergically co-operate in order to provide a fine tuning of protein expressions. In the present study, we designed a mathematically well-founded procedure to investigate the mutual positioning of transcription factors binding sites related to a given couple of transcription factors in order to evaluate the possible association between them. We obtained a list of highly related transcription factors couples, whose binding site occurrences significantly group together for a given set of gene promoters, identifying the biological contexts in which the couples are involved in and the processes they should contribute to regulate.

Key words: transcription factors, gene regulation, biological process, computational biology

1. Introduction

The regulation of gene expression is a fundamental mechanism driving biological processes. Transcriptional regulation rules the access of polymerase complex to the gene, activating or repressing the transcription process. Transcription Factors (TF) play a central role in this context; they are proteins able to bind specific DNA regions recognizing a short sequence of nucleotides called Transcription Factor Binding Sites (TFBS). There is a vast literature concerning TFs, starting from seminal and general papers[1,2] to the most specific ranging from the study of the mechanism that allows TFs to efficiently and rapidly find the target along the DNA helix,[3,4] to the study of the roles that specific TFs play in given biological tasks and how they influence the regulation of the transcription of their target genes.[5] TFBS occurrences along a DNA sequence can be statistically determined via a Position-specific Weight Matrix (PWM), that is essentially a matrix reporting the frequency of nucleotides for each position of the experimentally found binding sites. Several databases based on experimental evidences were designed collecting information about TFs and the corresponding TFBSs, TRANSFAC[6,7] and JASPAR[8] are often used as a reference for Eukariotes. In the review[9] different algorithmic approaches to predict TFBSs, starting from PWMs, are presented and compared. Despite the large number of genes in higher animals (~30,000 genes) the number of known TFs is one order of magnitude smaller, indicating a necessary interplay between different TFs in order to regulate the activity of all genes. TFBSs are often clustered in peculiar families and patterns made of several TFBSs can be found in the promoter region, i.e. thousands bases upstream the Transcription Starting Site (TSS), or also very far from the TSS in the Cis-regulatory modules (CRM). The interested reader can find in Ref. 10 (see also references therein) a comprehensive review of the update knowledge about CRM. In this work, we

will focus on the synergism of different TFs in the promoter region, trying to identify those TFs couples able to co-operate, and to discover the context in which those couples are involved in. The synergism between TFs in human DNA has been investigated by several studies focused on the analysis of mutual positioning of TFBSs along the DNA regions. Some studies, taking into consideration cross-species sequence comparison trying to identify possible regulatory modules in human DNA.[11–13] Other works focus on a specific class of TFs[14] or on a single TF[15] trying to identify the regulatory modules to which the considered TF belongs to. Finally in Ref. 16, the whole human genome has been investigated in order to reveal the couples of TFs that synergically act in the transcription regulation. In the present work, we developed a new computational method able to predict the possible interaction between couples of TFs analysing the statistical properties of the distribution of couples of TFBSs in the promoter region of human DNA, and comparing them with a random distribution of TFBSs. We obtained a list of highly related TFs couples, whose TFBSs occurrences significantly group together for a given set of promoter genes, identifying the biological context in which the couples are involved in and the process they should contribute to regulate.

2. Materials and methods

2.1. Extracting TFBSs of human genes

The list of all human genes was extracted from the Genome Browser (UCSC)[17] and all the associated 104,178 transcript sequences (genome version hg38) were retrieved by a proper R library[18] querying the UCSC data repository. For each transcript a promoter region, made of a sequence of 2,000 bp upstream the TSS, was extracted. In order to remove redundant information, the set of all promoter sequences was processed filtering out redundant overlapped transcripts, reducing their number (and the related promoter sequences) to 36,830. In this study, we used a set of 194 most studied and experimentally validated Human Transcription Factors according to MAPPER.[19] We associated to each TF the corresponding Position-specific Weight Matrix (PWM), i.e. the matrix reporting the frequency of nucleotides for each position of the experimentally validated binding sites. It is worth noting that a TF can be associated to different PWMs and, on the other hand, a PWM can be associated to more than one TF. A total number of 192 PWMs was considered in this work and each of them was associated in a biunivocal relation to one TF with only a few exceptions. For the sake of simplicity and readability, we refer to PWM models as TFs and vice versa. The PWMs matrices were retrieved from an open access repository available at the following address http://cistrome.org/~jian/motif_collection/databases/Transfac/pwm/ (12 October 2017, date last accessed). We applied the matchPWM() function, a sequence-based transcription factor binding site search algorithm, integrated into the Biostrings R library,[20] to extract and collect the positions list of all binding sites for both all considered transcripts and PWMs. For our purposes, we have considered an acceptance threshold probability of 0.9. The described procedure is clearly summarized in the workflow in Fig. 1.

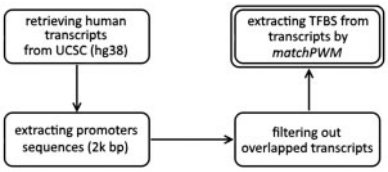

Figure 1. Workflow for the extraction of human TFBSs.

2.2. Transcription factor co-localization and deviation from randomness

In this Subsection, we present the indicator we have developed in order to determine whether two different TFs are linked with respect to a given promoter. Let $P = \{x_1 x_2 \dots x_L \ | x_i \in \{A, C, G, T\}$ be the promoter region of a given transcript, i.e. the set of $L = 2{,}000$ bases upstream the TSS of that transcript. Using PWM_i and PWM_j, i.e. the matrices associated to two different transcription factors TF_i and TF_j, we select, in the promoter P, all the TFBSs for TF_i and TF_j (see the previous Subsection for details) and we collect them in two sets, $\{TFBS_i\}$ and $\{TFBS_j\}$, respectively. The idea is to compare the number of couples of elements from $\{TFBS_i\}$ and $\{TFBS_j\}$ whose distance is closer than ℓ—where we set $\ell = 80$ bp—with the expected number of couples obtained by setting at random the occurrences of $\{TFBS_i\}$ and $\{TFBS_j\}$. As first we consider the case of random distribution of elements in $\{TFBS_i\}$ and $\{TFBS_j\}$ and we compute the expected number of joined occurrences within a distance ℓ. If there is just one binding site for factors TF_i and TF_j (i.e. there is just an element $x \in \{TFBS_i\}$ and $y \in \{TFBS_j\}$) put at random in a promoter of length L, the probability that their distance is exactly ℓ is

$$P(|x - y| = \ell) = 2\frac{L - \ell}{L(L-1)} \tag{1}$$

A simple computation gives the probability that their distance is less than or equal to ℓ

$$P(|x - y| \leq \ell) = \frac{(2L-1)\ell - \ell^2}{L(L-1)} \tag{2}$$

Going further, the probability that one element in $\{TFBS_i\}$ and several element $\{TFBS_j\}$, say n_j, put at random in a promoter of length L, have a distance less than or equal to ℓ is well-approximated by considering the probability of a single matching $P(|x - y| \leq \ell) = p$ as a Bernoulli event, and the probability of having k elements of $\{TFBS_j\}$ in an interval of size ℓ around the element $x \in \{TFBS_i\}$ is

$$P(k; |x - y| \leq \ell) = \binom{n_j}{k} p^k (1-p)^{n_j - k} \tag{3}$$

Finally, if there is more than one element in $\{TFBS_i\}$, say n_i, the expected number of occurrences of couple of elements in $\{TFBS_i\}$ and $\{TFBS_j\}$ within a distance ℓ can be approximated by

$$\begin{aligned} n(TF_i, TF_j) &= n_i \sum_{k=0}^{n_j} kP\left(k; \left|TFBS_i - TFBS_j\right| \leq \ell\right) = \\ &= n_i n_j p = n_i n_j \frac{(2L-1)\ell - \ell^2}{L(L-1)} \end{aligned} \tag{4}$$

while its variance is

$$\begin{aligned} \sigma^2(TF_i, TF_j) &= n_i \left[\sum_{k=0}^{n_i} k^2 P\left(k; \left|TFBS_i - TFBS_j\right| \leq \ell\right) - n(TF_i, TF_j)^2\right] \\ &= n_i n_j p(1-p) = n_i n_j \frac{\left[(2L-1)\ell - \ell^2\right]\left[(L-\ell)^2 - (L-\ell)\right]}{L^2(L-1)^2} \end{aligned} \tag{5}$$

The Bernoulli approximations in Equations (4) and (5) are valid when n_i and n_j are small with respect to ℓ and ℓ is small with respect

to L. By computing the actual number of couples (x, y), say $\nu(\mathrm{TF}_i, \mathrm{TF}_j)$, with $x \in \{\mathrm{TFBS}_i\}$ and $y \in \{\mathrm{TFBS}_j\}$ such that x and y are closer than ℓ, and using Equations (4) and (5) we build an indicator evaluating the deviation from randomness of the actual number of join occurrences

$$z(TF_i, TF_j) = \frac{\nu(TF_i, TF_j) - n(TF_i, TF_j)}{\sigma(TF_i, TF_j)} \qquad (6)$$

The above z-score function, measuring how many standard deviations the actual number of couples is far from the expected number of couples in the case of a random distribution, provides a reasonable value of the association between TFBS_i and TFBS_j. Significantly, high values of $z(\mathrm{TF}_i, \mathrm{TF}_j)$ suggest that the two considered TFs are associated, i.e. their TFBSs have a probability to occur as neighbours higher than the one expected for a random distribution. On the contrary when $z(\mathrm{TF}_i, \mathrm{TF}_j)$ is negative, with a high absolute value, elements in the two TFBSs occur as neighbours less frequently than expected in a random distribution, so that in this case the TFs are disassociated. In the following, in order to identify the synergy between transcription factors, we will focus on the case of high positive values of $z(\mathrm{TF}_i, \mathrm{TF}_j)$.

2.3. Computing similarity score of PWMs

The prediction algorithm of TFBSs is based on the PWMs obtained by experimental analysis. Since we are interested in identifying couples of TFs whose TFBSs occur together in a window of a given size, we have to be sure that the two corresponding models are enough different since if two models are very similar they are likely to share binding sites or a significant portion of them. We used two different approaches to evaluate models similarity. The former is a direct approach aiming to provide a similarity pairwise score of two models based on the comparison of nucleotide frequencies of the PWMs. The latter approach can be defined as 'a posteriori', since we evaluate the closeness of models by computing the number of overlapped binding sites with respect to the total number of couples whose distance is closer than ℓ. Using a linear combination of the above described similarity scores we determine if two models are 'structurally' far from each other and we can significantly evaluate the association between two truly different models according to the algorithm described in the previous section.

2.4. PWMs distance based on Jensen-Shannon divergence

The Jensen-Shannon divergence (JS) is a symmetrized and smoothed version of the Kullback-Leibler divergence and it is often used to tell how two frequency distributions are close to each other.[21] Here, we use Jensen-Shannon divergence to identify similar PWMs using the nucleotide frequencies as probability distribution. The Jensen-Shannon divergence of two frequency distributions, P and Q, can be computed using the Shannon entropy according to the following equation:

$$\mathrm{JS}(P, Q) = H\left(\frac{1}{2}P + \frac{1}{2}Q\right) - \frac{1}{2}H(P) - \frac{1}{2}H(Q) \qquad (7)$$

where $H(X)$ is the Shannon entropy of the distribution X

$$H(X) = -\sum_{k=1}^{N} x_k \log(x_k)$$

Given a couple of models, PWM_1 and PWM_2, we indicate with P^i_k and Q^j_k the frequency of the nucleotide k (the index k ranges from 1 to 4 indicating the four nucleotides A, C, G, T) in the position i of the model PWM_1 (the index i ranges from 1 to I, where I is the length of the model PWM_1) and position j of the model PWM_2 (the index j ranges from 1 to J, where J is the length of the model PWM_2), respectively. For example, P^1_4 is the frequency of the nucleotide T in position 1 of the first model while Q^2_3 is the frequency of the nucleotide G in position 2 of the second model. The Jensen-Shannon divergence of the nucleotide distributions at position i of the first model and at position j of the second model is defined as

$$\mathrm{JS}(P, Q; i, j) = -\sum_{k=1}^{4} \frac{P^i_k + Q^j_k}{2} \log \frac{P^i_k + Q^j_k}{2} + \frac{1}{2}\sum_{k=1}^{4} P^i_k \log P^i_k + \sum_{k=1}^{4} Q^j_k \log Q^j_k \qquad (8)$$

We measure the similarity of two models with respect to a given alignment by summing up the Jensen-Shannon divergences as in Equation (8)) of each couple of distributions corresponding to the aligned positions (empty boxes in the overlapping area, see Fig. 2) (i, $i+\mathrm{a}$), where a is the offset between the starting positions of model 1 and 2 in that alignment, plus a penalty score for the non overlapped positions ("p" marked boxes in the penalty area, see Fig. 2):

$$\mathrm{JS}(P, Q; a) = \sum_{i \in I} \mathrm{JS}(P, Q; i, i + a) + K n_o \qquad (9)$$

where I is the set of overlapped positions in the given alignment and n_o is the number of non overlapped positions of the shortest model. The penalty K for each of non overlapped positions is determined by computing the probability distribution of the Jensen-Shannon divergence in the case of two random nucleotide distributions. The value $K \simeq 0.23$ is obtained by summing up 2 S.D. to the mean value of the random-nucleotide Jensen-Shannon divergence distributions. Finally, the model similarity score of two models JSD(P, Q) is defined as the smallest score JS(P, Q; a) out of all the possible alignments:

$$\mathrm{JSD}(P, Q) = \min_{a \in \mathrm{AL}} \{\mathrm{JS}(P, Q; a)\} \qquad (10)$$

where AL is the set of all possible alignments with a non-empty superposition. The need of a 'strong'' non-overlapping penalty comes out in order to select alignment with a significant overlapping between the two models (i.e. to avoid bias due to a good affinity between too small part of the two models).

2.5. PWMs distance based on TFBS overlapping

Here, we introduce a different similarity function between pairs of TFs based on the percentage of overlapped binding sites they share. First of all, we consider two TFBSs to be overlapped if their distance is smaller than the half of the size of the smaller PWM, i.e. $|\mathrm{TFBS}_i - \mathrm{TFBS}_j| < \frac{\min\{\mathrm{PWM}_1,\ \mathrm{PWM}_2\}}{2}$. If the TFBSs of two TSs are often overlapped it is likely that they result to be associated, according to our algorithm, showing a high Z-score. They are indeed often close because the TFs recognize the same binding sites (or a large portion of them). Given a couple of TFs, TF_i and TF_j, the percentage of all overlapping occurrences of TFBSs with respect to all

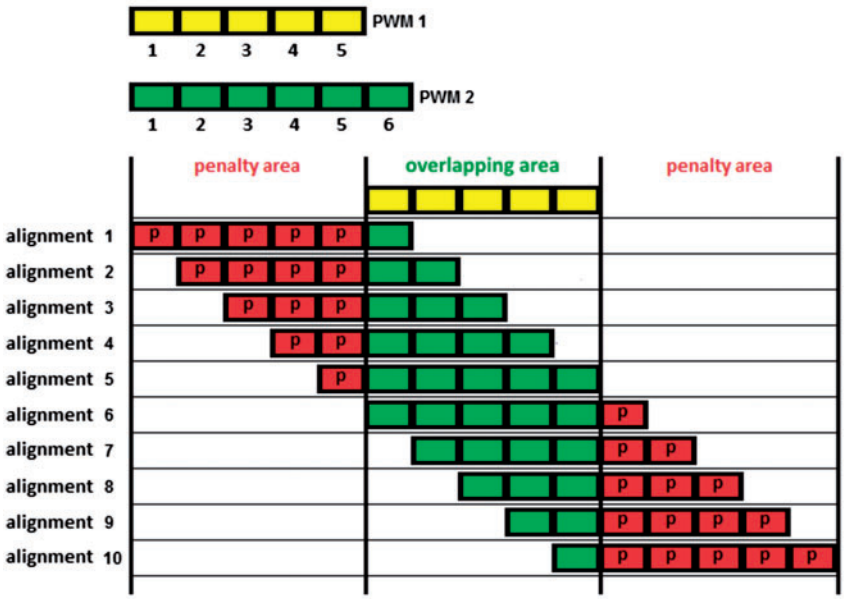

Figure 2. Scheme of the algorithm for computing similarity score of models.

the joined occurrences in a radius ℓ in all the transcript promoter sequences, namely OVD(TF_i, TF_j), is defined as follows:

$$\mathrm{OVD}(\mathrm{TF}_i, \mathrm{TF}_j) = \frac{\mu_{\mathrm{Tot}}(\mathrm{TF}_i, \mathrm{TF}_j)}{\nu_{\mathrm{Tot}}(\mathrm{TF}_i, \mathrm{TF}_j)} \qquad (11)$$

where $\nu_{\mathrm{Tot}}(\mathrm{TF}_i, \mathrm{TF}_j)$ is the total number of the TFBSs couples (one from TF_i and the other from TF_j) falling together in the considered radius of size ℓ and $\mu_{\mathrm{Tot}}(\mathrm{TF}_i, \mathrm{TF}_j)$ is the total number of TFBSs overlapped couples (one from TF_i and the other from TF_j), for all the transcript promoter sequences. In the next section (as reported in Fig. 3), OVD scores are plotted against JSD scores in order to select significant couples.

2.6. Protein-protein interaction network analysis

In order to validate the synergy of TFs couples identified by the algorithm here presented we analysed Protein-Protein Interaction network (PPI) to test the interaction of this couples in the network by using shortest path.

Protein-Protein Interaction network was downloaded from STRING database.[22] We considered all the possible interaction sources between proteins giving a score threshold of 0.7 (high confidence interaction). The obtained network is made of 719,288 edges (interactions) and 14,932 nodes (proteins). The iGraph R package (https://cran.r-project.org/web/packages/igraph/index.html (12 October 2017, date last accessed)) was used to compute the shortest path from the adjacency matrix of the graph.

3. Results and discussion

3.1. Selecting significant TF couples

In the previous Section, Equation (6) has been derived in order to provide a measure of association between each couple of transcription factors (TF_i, TF_j) with respect to a given transcript TR_k. Our goal is to find out couples of TFs that are significantly related in gene-regulation mechanism. As previously mentioned, the higher the value of $z(\mathrm{TF}_i, \mathrm{TF}_j; \mathrm{TR}_k)$, the higher the association between TF_i and TF_j in the transcript promoter TR_k sequence will be. Since for each transcript, we have to test a considerably large number of couples (TF_i, TF_j) and we test the association between couples in a large number of transcripts, we select only those couples for which the value of $z(\mathrm{TF}_i, \mathrm{TF}_j; \mathrm{TR}_k)$ is higher than 5 (meaning that the number of occurrence of couples TF_i TF_j in a radius ℓ is far from that expected by chance more than 5 S.D.) and we indicate with $NZ_5(\mathrm{TF}_i, \mathrm{TF}_j)$ the number of transcripts TR_k for which the couple (TF_i, TF_j) has a z-score greater than 5. Given a couple (TF_i, TF_j) it is reasonable to expect that a large value of $NZ_5(\mathrm{TF}_i, \mathrm{TF}_j)$ indicates that the two models TF_i, TF_j are functionally associated. However, a high value of $NZ_5(\mathrm{TF}_i, \mathrm{TF}_j)$ could refer to a couple of models with a high similarity score rather than a functional association between TF_i and TF_j. Therefore, to correctly identify relevant associated transcription

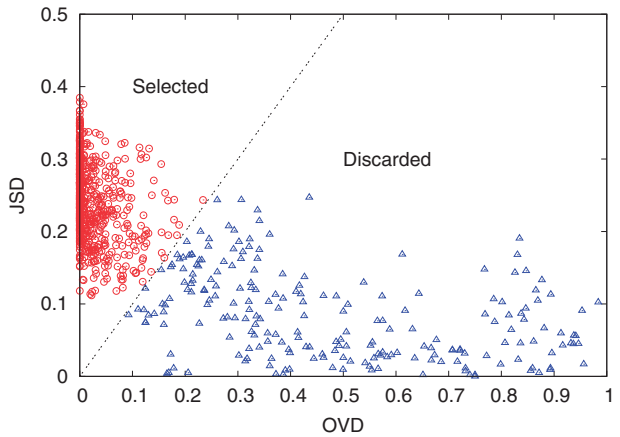

Figure 3. Each point represents a TF couple whose Z-scores are significant ($Z > 5$) for at least 20 transcripts. The plane is divided into two regions: the upper-left sector, where Overlapping Score is higher than Similarity Score, provides selected couples. Lower-right sector, where Overlapping Score is lower than Similarity Score, provides Discarded couples.

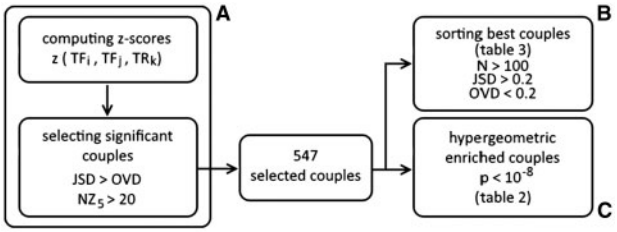

Figure 4. Workflow for the analysis of TF couples. In panel A, the procedure for the selection of relevant couples is depicted. In panel B, the analysis of significant couples is described both in terms of enrichment tests and sorting of best couples.

Table 1. Categories and terms for the enrichment analysis

Category	Number of terms	Number of genes
OMIM expanded	187	2,178
Tissue protein expression from proteomicsDB	207	62,307
KEGG 2015	179	3,800
OMIM disease	90	1,759
GO molecular function	1,136	12,753
GO cellular component	641	13,236
GO biological process	5,192	14,264
Chromosome location	386	32,740

factors, we will use the overlapping score OVD(TF_i, TF_j) and the similarity score JSD(TF_i, TF_j). In Fig. 3, OVD(TF_i, TF_j) and JSD(TF_i, TF_j) values are shown for all the TF couples which have $NZ_5(TF_i, TF_j) > 20$. Since the higher the value of JSD the lower the similarity between two models is, while the higher the values of OVD the higher the similarity between binding sites is, as selection criterion to identify significant couples we used the inequality JSD > OVD. In such a way, the plane (OVD, JSD) is divided into two regions: the upper-left sector, where models are not similar and the overlapping is low (selected couples), while in the lower-right sector, where models are similar and the overlapping is high (discarded couples). It is worth remarking that even if measures JSD and OVD could appear redundant, Fig. 3 clearly stated that the two measures are quite different: there are couples with similar JSD and very different OVD and vice versa. In fact JSD measure only depends on the related PWM while OVD measure depends on the considered dataset, e.g. two models can bind the same sequence even if they are different. Considering a combination of both measures we obtain a more robust selection criterion.

In Fig. 4 (panel A) is sketched the procedure we use to select the 547 couples identifying pairs of transcription factors that are likely to synergistically regulate gene transcription. In the following sections, we will analyse the set of selected couples in order to associate to them a biological task and to find the most relevant associated couples according to their statistical features (see Fig. 4 panel B).

3.2. Enrichment analysis and validation

In order to evaluate the biological significance of the obtained TF couples, we performed hypergeometric enrichment tests.[23] Given a couple of models, we want to test whether the set of genes potentially regulated by the synergy of the corresponding two TFs, is enriched in a particular category-term. We wondered whether those genes are linked in a given biological task in order to associate the TF couple to a given biological process. We selected a list of categories and related terms, reported in Table 1, that identify different biological tasks and contexts. We then performed for each identified TF couple hypergeometric enrichment tests related to all the considered terms.

For each TF couple we obtained a list of *P*-values linked to the corresponding terms. The enrichment results were then compared with those of ten random sets of genes. Each set was generated preserving the same number of sample couples (i.e. 547) and the same number of genes for each couple. The results are reported in Fig. 5 where solid plot represents the enrichment of couples identified by our algorithm and dashed plots show enrichment of random sets. Each point of the plots shows in log-scale the number of TF couples (*y*-axis) with at least one term whose *P*-value falls in the corresponding *P*-value interval (*x*-axis). For example in

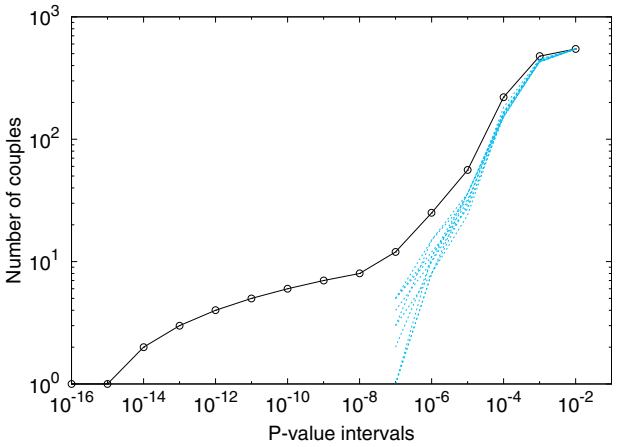

Figure 5. Number of couples enriched in at least one term as a function of their smallest *P*-value. Solid plot shows the enrichment of the couples identified by our algorithm, dashed plots show the enrichment of random couples for comparison.

the black plot, there is only one couple (M00106—M00967) having a term [ubiquitinyl hydrolase activity (GO: 0036459)] enriched with a *P*-value smaller than 10^{-18} (3.63E–19), another couple (M00739—M00799) for *P*-value smaller than 10^{-14} and three different couples for *P*-value smaller than 10^{-13} and so on (see Table 2 where all the enriched terms are ordered according to their *P*-values). It is worth noting that some couples have more than one term with a significant *P*-value, for example couple (M00106—M00967) shows several terms with a remarkable enrichment: GO Molecular function—ubiquitinyl hydrolase activity (GO: 0036459)—*P*-value 3.63E–19, Go Molecular function—cysteine-type peptidase activity (GO: 0008234)—*P*-value 4.56E–17, Go Biological Process—Ubiquitin-dependent protein catabolic process (GO: 0006511), and so on. The comparison of enrichment *P*-values, related to couples identified by our algorithm (black plot), with those related to random sets (ten coloured plots) clearly highlights the significance and consistency of our analysis, since none of the random sets shows *P*-values smaller than 10^{-7}. Also analysing the plots for larger *P*-values intervals the number of enriched couples is significantly higher than all the corresponding numbers of random sets (taking into account the log-scale units in *y*-axis).

3.3. Analysis of relevant transcription factor couples

In this section, we analyse the most significant couples identified by our algorithm. Among the 547 selected couples, we consider as particularly relevant those couples with the smallest enrichment *P*-value or

Table 2. The PWM model IDs and the corresponding TF names are reported in columns 1–4 for each couple; the category and the related term the couple resulted enriched in are reported in columns 5–6; column 7 reports the related *P*-value of the hypergeometric test

Model 1	TF 1	Model 2	TF 2	Category	Term	P-value
M00106	CDP CR3+HD	M00967	HNF4 COUP	GO Molecular function	ubiquitinyl hydrolase activity (GO: 0036459)	3, 63E–19
M00106	CDP CR3+HD	M00967	HNF4 COUP	GO Molecular function	cysteine-type peptidase activity (GO: 0008234)	4, 56E–17
M00739	E2F-4: DP-2	M00799	Myc	GO Biological process	Gonadal mesoderm development (GO: 0007506)	6, 60E–15
M00106	CDP CR3+HD	M00967	HNF4 COUP	GO Biological process	Ubiquitin-dependent protein catabolic process (GO: 0006511)	3, 59E–14
M00736	E2F-1: DP-1	M00799	Myc	GO Biological process	Gonadal mesoderm development (GO: 0007506)	4, 09E–14
M00106	CDP CR3+HD	M00967	HNF4 COUP	GO Biological process	Modification-dependent protein catabolic process (GO: 0019941)	4, 42E–14
M00106	CDP CR3+HD	M00967	HNF4 COUP	GO Biological process	Modification-dependent macromolecule catabolic process (GO: 0043632)	4, 90E–14
M00106	CDP CR3+HD	M00967	HNF4 COUP	GO Biological process	Proteolysis involved in cellular protein catabolic process (GO: 0051603)	8, 43E–14
M00736	E2F-1: DP-1	M00799	Myc	GO Biological process	Nucleosome organization (GO: 0034728)	1, 42E–13
M00777	STAT	M00980	TBP	Chromosome location	Chr5q13	1, 54E–13
M00736	E2F-1: DP-1	M00799	Myc	GO Biological process	Protein-DNA complex subunit organization (GO: 0071824)	4, 01E–13
M00736	E2F-1: DP-1	M00799	Myc	GO Biological process	Nucleosome assembly (GO: 0006334)	2, 81E–12
M00457	STAT5A	M00980	TBP	Chromosome location	Chr5q13	6, 40E–12
M00736	E2F-1: DP-1	M00799	Myc	GO Biological process	Protein-DNA complex assembly (GO: 0065004)	8, 69E–12
M00223	STATx	M00980	TBP	Chromosome location	Chr5q13	4, 82E–11
M00799	Myc	M00927	AP-4	GO Biological process	Gonadal mesoderm development (GO: 0007506)	1, 12E–10
M00739	E2F-4: DP-2	M00799	Myc	GO Biological process	Mesoderm development (GO: 0007498)	2, 71E–10
M00739	E2F-4: DP-2	M00799	Myc	GO Biological process	Mesenchyme development (GO: 0060485)	4, 02E–10
M00736	E2F-1: DP-1	M00799	Myc	GO Biological process	Mesoderm development (GO: 0007498)	1, 66E–09
M00736	E2F-1: DP-1	M00799	Myc	GO Biological process	Mesenchyme development (GO: 0060485)	2, 46E–09
M00739	E2F-4: DP-2	M00799	Myc	GO Biological process	Nucleosome assembly (GO: 0006334)	4, 46E–09
M00462	GATA-6	M00921	GR	GO Molecular function	FK506 binding (GO: 0005528)	7, 13E–09
M00462	GATA-6	M00921	GR	GO Molecular function	Macrolide binding (GO: 0005527)	7, 13E–09
M00739	E2F-4: DP-2	M00799	Myc	GO Biological process	Protein-DNA complex assembly (GO: 0065004)	9, 90E–09
M00739	E2F-4: DP-2	M00799	Myc	GO Biological process	Nucleosome organization (GO: 0034728)	1, 44E–08
M00655	PEA3	M00803	E2F	Chromosome location	Chr5q13	1, 50E–08
M00799	Myc	M00803	E2F	GO Biological process	Protein autophosphorylation (GO: 0046777)	2, 05E–08
M00739	E2F-4: DP-2	M00799	Myc	GO Biological process	Protein-DNA complex subunit organization (GO: 0071824)	2, 73E–08
M00059	YY1	M00148	SRY	GO Molecular function	Sodium ion transmembrane transporter activity (GO: 0015081)	2, 96E–08
M00415	AREB6	M00706	TFII-I	KEGG	Valine leucine and isoleucine biosynthesis	4, 10E–08

with the best associated parameters (highest number of transcripts, smallest OVD and highest JSD) as reported in Fig. 4 panel B. The list of couples with the smallest *P*-values (smaller than 10^{-7}) is reported in Table 2. The analysis of the couples reported in the table suggests several observations, providing and/or confirming evidence of the synergism of several TF couples and the related biological contexts they are involved in. The smallest *P*-values are reported for the couple CDP (also known as Cutl1) (M00106)–HNF4 (M00967), as already mentioned above. The related enriched terms are mainly associated (4 out of 6) with catabolic processes while the other two terms are ubiquitinyl hydrolase activity and cysteine-type peptidase activity. Concerning the couple Myc (M00799)–E2F (M00803) (and its cluster E2F–1 and E2F–4) it can be observed that it appears several times in the table associated to different terms. These terms can be grouped into main activities/contexts regarding: (1) gonadal mesoderm development and mesenchyme development, (2) Nucleosome assembly and organization, Protein-DNA complex assembly and Protein-DNA Complex Subunit Organization and (3) Protein Autophosphorylation. It is worth noting that in literature there are many references associated to the couple Myc–E2F. In Ref. 24, it has been shown how the two TFs are linked via mSin3A, a core component of a large multiprotein co-repressor complex with associated histone deacetylase (HDAC) enzymatic activity. As reported in Ref. 25, it has been identified a subunit of the complex NuA4 as the product of TRA1, an ATM-related gene homologous to human TRRAP, an essential cofactor for c-Myc- and E2F-mediated oncogenic transformation. A further link between Myc and E2F was studied in Ref. 26 highlighting their relationships with chromatin structure and stability. It is remarkable and can be source of further investigations the association of the couple AREB (M00415)–TFII-I (M00706) (also known as GTF2I) with KEGG pathway Valine leucine and isoleucine biosynthesis (the couple showing a *P*-value of 4.10E–8). Interestingly, transcription factors of STAT family (STAT—M00777, STATx—M00223 and STAT5A–M00457) strongly interact with TBP (M00980) in the promoters of genes belonging to the region of chr5q13, that is known to be associated with various neurological disorders and pathologies such as Spinal Muscular Atrophy, Hairy Cell Leukemia and it is also connected to Alcohol Dependence.[27–29] Finally, it could deserve attention the couple Areb6 (M00414)–E2F (M00803) associated to the term Mental retardation, even if not reported in Table 2, it shows a remarkable *P*-value ($6{\cdot}10^{-6}$). Areb6 belongs to the Zeb transcription factor family that has been shown to be involved in mental retardation syndromes.[30] In Table 3, the best couples, in terms of association, are reported ordered by the number

Table 3. The PWM model IDs and the corresponding TF names are reported in columns 1–4 for each couple; Column 5 reports the number of transcripts whose *Z*-scores, related to the couple, are higher than 5; the average *Z*-score of all the significant ($Z > 5$) transcripts is reported in column 6; Overlapping and Similarity Scores are reported in columns 7 and 8, respectively

Model 1	TF 1	Model 2	TF 2	Number of transcripts	Average *Z*-score	Overlapping score	Similarity score
M00083	MZF1	M00649	MAZ	1,014	7, 77	0, 19	0, 21
M00083	MZF1	M00803	E2F	456	6, 77	0, 09	0, 21
M00803	E2F	M00976	AhR Arnt HIF-1	338	6, 63	0, 03	0, 21
M00706	TFII-I	M00971	Ets	317	7, 71	0, 16	0, 24
M00799	**Myc**	**M00803**	**E2F**	**293**	**7, 16**	**0, 00**	**0, 23**
M00706	TFII-I	M00803	E2F	275	6, 61	0, 00	0, 27
M00148	SRY	M00747	IRF-1	254	7, 77	0, 00	0, 20
M00148	SRY	M00471	TBP	239	6, 78	0, 02	0, 21
M00148	SRY	M00980	TBP	203	6, 45	0, 00	0, 23
M00698	HEB	M00803	E2F	196	6, 35	0, 02	0, 29
M00649	MAZ	M00658	PU.1	187	7, 49	0, 03	0, 22
M00649	MAZ	M00799	Myc	184	6, 85	0, 00	0, 33
M00799	Myc	M00933	Sp1	182	7, 06	0, 01	0, 28
M00462	GATA-6	M00471	TBP	182	6, 47	0, 08	0, 21
M00799	Myc	M00931	Sp1	167	6, 97	0, 00	0, 30
M00803	E2F	M00927	AP-4	160	6, 29	0, 01	0, 24
M00801	CREB	M00803	E2F	153	6, 43	0, 00	0, 28
M00706	TFII-I	M00931	Sp1	141	6, 58	0, 04	0, 22
M00649	MAZ	M00971	Ets	132	7, 32	0, 00	0, 23
M00933	Sp1	M00976	AhR Arnt HIF-1	127	6, 65	0, 02	0, 21
M00803	E2F	M00981	CREB ATF	127	6, 72	0, 00	0, 23
M00799	Myc	M00932	Sp1	121	7, 17	0, 00	0, 28
M00148	SRY	M00706	TFII-I	120	8, 58	0, 00	0, 31
M00931	Sp1	M00976	AhR Arnt HIF-1	117	6, 47	0, 01	0, 20
M00803	E2F	M00917	CREB	115	6, 81	0, 00	0, 25
M00008	Sp1	M00706	TFII-I	114	6, 36	0, 02	0, 21
M00791	HNF3	M00975	RFX	107	6, 58	0, 00	0, 20
M00471	TBP	M00747	IRF-1	107	6, 32	0, 02	0, 25
M00649	MAZ	M00976	AhR Arnt HIF-1	106	6, 28	0, 00	0, 26
M00148	SRY	M00962	AR	106	6, 04	0, 00	0, 20
M00148	SRY	M00789	GATA	106	6, 08	0, 00	0, 23
M00148	SRY	M00975	RFX	104	6, 20	0, 00	0, 21
M00008	Sp1	M00799	Myc	102	6, 77	0, 00	0, 29
M00775	NF-Y	M00803	E2F	101	6, 24	0, 03	0, 20

The couple **Myc, E2F** is reported in bold to highlight it is also included in Table 2.

Table 4. Frequencies of couples as a function of the shortest path (SP) distance for three classes of protein couples: (i) the selected 547 TF couples (namely BEST, first row), (ii) all the couples of TFs (namely ALL, second row) and iii) 10 random sample sets made of 547 randomly picked protein couples from the whole PPI (namely RANDOM, mean and standard deviation in the third and fourth row, respectively)

	SP 1	SP 2	SP 3	SP 4	SP 5	SP 6	SP 7
BEST	0.117318	0.478585	0.284916	0.10987	0.009311	0	0
ALL	0.056789	0.463694	0.376242	0.084080	0.014167	0.004661	0.000368
RANDOM (mean)	0.003291	0.065601	0.377048	0.385820	0.134512	0.027972	0.005346
RANDOM (stdv)	0.000943	0.006363	0.016408	0.008298	0.009771	0.003329	0.002414

of significant Transcripts. We selected those couples with a number of significant transcripts (*Z*-score higher than 5) higher than 100, with an Overlapping score (OVD) smaller than 0.2 and a Similarity score (JSD) higher than 0.2 (it is also reported in the fifth column the average *Z*-Score). The couple MZF1 (M00083)—MAZ (M00649) shows a *Z*-score higher than 5 in 1,014 transcript promoter sequences (with an average *z*-score of 7.77—more than 7 S.D. far from the expected average value). Myc-associated zinc finger protein (MAZ) and Myeloid zinc finger 1 (MZF1) are both transcription factors characterized by a zinc finger small protein structural motif. The associated PWMs show a similar common sub-motif characterized by a GGGA sequence. Nevertheless the PWMs have different length (6 and 8) and the correspondent similarity score, JSD, results to be 0.21 meaning that the two PWMs are similar but globally not so close to each other. Significantly, the OVD overlapping score is quite low 0.188 meaning that, on average, among all the identified TFBSs couple (both in a window of 80 bp) of the two TFs, less than 1 out of 5 couples are overlapped. This is why we included this couple in the selected set of TFs couples. The surprisingly high number of transcripts for which the two TFs co-occur should deserve deeper investigations. The couple MZF1 (M00083)–E2F (M00803) shows 456 transcripts with a *Z*-score higher than 5 (with an average *z*-score of 6, 77, OVD = 0.09 and JSD = 0.21). Interestingly they are involved in several diseases, in particular, as reported in Ref. 31, they are both potential key regulators of PKD1 and PKD2 whose mutations are linked with autosomal dominant polycystic kidney disease (ADPKD). We found one transcript of PKD1 uc002cos.1 with a *Z*-score for the couple equal to 2 and two transcripts of PKD2 uc003hre.3 and uc011cdg.2 with *Z*-score equal to 5.80 and 3.42, respectively. It is interesting that a minor groove binding protein SRY (M00148) is associated to both the models of TBP (M00471 and M00980) with a number of significant transcripts equal to 239 and 203, respectively, a very small OVD close to 0 for both and JSD 0.21 and 0.23, respectively. It is worth noting that the couple Myc (M00799)–E2F (M00803) (also included in Table 2) shows a number of significant transcripts equal to 293, meaning that, besides a clear association with given biological contexts, as discussed in the previous section, there is a strong synergism between the two TFs confirmed by the values reported in the table. Concerning the couple Maz (M00649)–Pu.1 (M00658) (number of significant transcripts equal to 187) it has been shown that three transcription factors Maz, PU.1 and ARNT show significant recognition elements among similarly up or down-regulated genes involved in hematopoietic differentiation or leukemogenesis.[32] We note that also the couple Maz (M00649)–ARNT (M00976) is included in the table with a number of significant transcripts equal to 106.

In order to further validate and provide significance to obtained results, we computed the shortest paths between couples of TFs related to the Protein-Protein Interaction (PPI) network downloaded from String database.[22] We found a significant overall difference between shortest path distribution of the set of couples selected by our method and the set of all TFs couples. In Table 4, we report the percentages of couples as a function of the Shortest Path (SP) distances for three classes of protein couples: (i) the selected 547 TF couples (namely BEST, first row), (ii) all the couples of TFs (namely ALL, second row) and (iii) 10 random sample sets made of 547 randomly picked protein couples from the whole PPI (namely RANDOM, mean and standard deviation in the third and fourth row, respectively). The majority of couples (around 50%) related to TFs shows a SP equal to 2 both for couples obtained by our algorithm and for all TF couples, while only around 6% of protein random couples has a SP distance equal to 2. The most relevant result is in the difference of the frequency for SP = 1 (indicating a direct interaction) that is ∼0.12 for our best couples and ∼0.057 for all TF couples (i.e. the ratio is greater than 2), and also for SP = 3 (∼0.28 our best couples versus ∼0.37 for all TF couples). The distribution for the random protein data shows, not surprisingly, a complete different pattern (the most part of couples having SP equal to 3 and 4). Even considering the limitations of this kind of analysis—a PPI is a global representation of potential interactions between proteins (and consequently TFs) that not always (referring to time and space) is an actual interaction and, moreover, the results depend on the threshold chosen to select significant interactions—it reveals that TFs couples selected by our method show an overall stronger relationship than those in the set of all TFs couples. In particular, we chose a quite strict threshold of 0.7 (as reported in the Material and methods section), so the significance of results has to be found in the ratio between the number of our selected couples and all TFs couples at SP 1 (ratio equal to 2) rather than in the percentages (12% and 6%, approximately) that could significantly change since they depend on the threshold.

4. Conclusions

A better understanding of the mechanisms driving the regulation of protein expression is an essential requisite to shed light on the behaviour of cells. Transcription factors play a central role in this extremely complex task and it has been shown that they synergically co-operate in order to provide a fine tuning of protein expressions. Among the different methods able to detect and identify TFs interplay, a very important resource is the computational methods to which our work mainly refers. In this work, we present a mathematically well-founded procedure able to identify TFs couples that act together, inferring for several of those couples the biological context they are involved in. We introduced a new and robust statistical method based on the use of a good Bernoulli approximation and on a *z*-score measures able to discriminate between random and non-random co-occurrences of couple of TFs. We

extended previous methods based on randomize sequences as in Refs. 15 and 16. In order to avoid biases due to the structural similarity of different models Overlapping and Similarity scores were designed to select TF couples that are significant for the analysis. We used such a method to find pairs of associated transcription factors in the set of all human promoter sequences (selected with a careful analysis from all human transcripts) and we also performed enrichment analysis on the set of genes regulated by identified couples. This analysis provides consistence to our results but also provides a biological context to be associated to the couples. Moreover, we also performed network analysis showing that TFs couples identified by the algorithm here presented are closer than expected in the protein-protein interaction network in terms of shortest path.

To our knowledge this is the only work concerning synergy of Transcription Factors taking into account all those features. Some of the couples emerging in this study are already known to be linked in several biological contexts (such as Myc–E2F,[24–26] while in other cases our results can lead to hypothesize links between TFs couples and diseases (as for the STAT family that strongly interact with TBM in the promoters of genes belonging to the region of chr5q13 involved in several diseases.[27–29] Finally, some other couples were not previously identified (such as the couple MZF1–MAZ) that would deserves further investigation.

According to our algorithm two TFs could be associated in different ways, for example they could be co-operative in a strict sense or concurrent in the sense that the presence of one of the two impedes the presence of the other. A small size window, as the one we use, leads us to hypothesize that two transcription factor proteins that could bind sites within the window, either are sufficiently close to each other to physically interact, or only one of the two is able to bind its TFBS because of the steric hindrance.

This work represents a step in the direction of designing complex gene regulatory networks, and it provides information on TFs association that could be useful in this context. The identification of significant TFs couples could be of help in the view of artificially altering the regulation of genes by inhibiting the interaction between given TFs couples.

Conflict of interest

None declared.

References

1. Mitchell, P.J. and Tjian, R. 1989, Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins, *Science*, **245**, 371–8.
2. Latchman, D.S. 1997, Transcription factors: an overview, *Int. J. Biochem. Cell Biol*., **29**, 1305–12.
3. Zhou, H.X. 2011, Rapid search for specific sites on DNA through conformational switch of nonspecifically bound proteins, *Proc. Nat. Acad. Sci. U.S.A*., **108**, 8651–6.
4. Berg, O.G., Winter, R.B. and Von Hippel, P.H. 1981, Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory, *Biochemistry*, **20**, 6929–48.
5. Pahl, H.L. 1999, Activators and target genes of Rel/NF-kB transcription factors, *Oncogene*, **18**, 6853–66.
6. Wingender, E., Chen, X., Hehl, R., et al. 2000, TRANSFAC: an integrated system for gene expression regulation, *Nucleic Acids Res*., **28**, 316–9.
7. Wingender, E. 2008, The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation, *Brief. Bioinform*., **9**, 326–32.
8. Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W. and Lenhard, B. 2004, JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic Acids Res*., **32**, D91–4.
9. Tompa, M., Li, N., Bailey, L.T., Church, G.M., et al. 2005, Assessing computational tools for the discovery of transcription factor binding sites, *Nat. Biotechnol*., **23**, 137–44.
10. Suryamohan, K. and Halfon, M.S. 2015, Identifying transcriptional cis-regulatory modules in animal genomes, *Wires. Dev. Biol*., **4**, 59–84.
11. Blanchette, M., Bataille, A.R., Chen, X., et al. 2006, Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome. Res*., **16**, 656–68.
12. Prabhakar, S., Poulin, F., Shoukry, M., et al. 2006, Close sequence comparisons are sufficient to identify human cis-regulatory elements, *Genome Res*., **16**, 855–63.
13. Hu, Z., Hu, B. and Collins, J.F. 2007, Prediction of synergistic transcription factors by function conservation, *Genome Biol*., **8**, R257.
14. Yu, X., Lin, J., Zack, D.J. and Qian, J. 2006, Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues, *Nucleic Acids Res*., **34**, 4925–36.
15. Nandi, S., Blais, A. and Ioshikhes, I. 2013, Identification of cis-regulatory modules in promoters of human genes exploiting mutual positioning of transcription factors, *Nucleic Acids Res*., **41**, 8822–41.
16. Hannenhalli, S. and Levy, S. 2002, Predicting transcription factor synergism, *Nucleic Acids Res*., **30**, 4278–84.
17. Kent, W.J., Sugnet, C.W., Furey, T.S., et al. 2002, The human genome browser at UCSC, *Genome Res*., **12**, 996–1006.
18. Gentleman, R.C., Carey, V.J., Bates, D.M., et al. 2004, Bioconductor: open software development for computational biology and bioinformatics, *Genome Biol*., **5**, 1.
19. Marinescu, V.D., Hohane, I.S. and Riva, A. 2005, The MAPPER database: a multi-genome catalog of putative transcription factor binding sites, *Nucleic Acids Res*., **33** (**suppl 1**), D91–7. [CVOCROSSCVO]
20. Pagès, H., Aboyoun, P., Gentleman, R. and DebRoy, S. 2009, String objects representing biological sequences, and matching algorithms. R package version 2.2.
21. Lin, J. 1991, Divergence measures based on the shannon entropy, *IEEE Trans. Inform. Theory*, **37**, 145–51.
22. Szklarczyk, D., Morris, J.H., Cook, H., et al. 2017, The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, *Nucleic Acids Res*., **45**, D362–8.
23. Lee, J.K., ed. 2010, *Statistical Bioinformatics for Biomedical and Life Science Researchers*. Wiley-Blackwell, Hoboken, New Jersey.
24. Dannenberg, J.H., David, G., Zhong, S., van der Torre, J., Wong, W.H. and Depinho, R.A. 2005, mSin3A corepressor regulates diverse transcriptional networks governing normal and neoplastic growth and survival, *Genes Dev*., **19**, 1581–95.
25. Allard, S., Utley, R.T., Savard, J., et al. 1999, NuA4, an essential transcription adaptor/histone H4 acetyltransferase complex containing Esa1p and the ATM-related cofactor Tra1p, *Embo J*., **18**, 5108–19.
26. Albert, T., Wells, J., Funk, J.O., et al. 2001, The chromatin structure of the dual c-myc promoter P1/P2 is regulated by separate elements, *J. Biol. Chem*., **276**, 20482–90.
27. Lin, P., Hartz, S.M., Wang, J.C., et al. 2012, Copy number variations in 6q14. 1 and 5q13. 2 are associated with alcohol dependence, *Alcohol. Clin. Exp. Res*., **36**, 1512–8.
28. Also-Rallo, E., Alias, L., Martinez-Hernandez, R., et al. 2011, Treatment of spinal muscular atrophy cells with drugs that upregulate SMN expression reveals inter-and intra-patient variability, *Eur. J. Hum. Genet*., **19**, 1059–65.

29. Wu, X., Ivanova, G., Merup, M., et al. 1999, Molecular analysis of the human chromosome 5q13. 3 region in patients with hairy cell leukemia and identification of tumor suppressor gene candidates, *Genomics*, **60**, 161–71.
30. Zweier, C., Albrecht, B., Mitulla, B., et al. 2002, 'Mowat-Wilson' syndrome with and without hirschsprung disease is a distinct, recognizable multiple congenital anomalies-mental retardation syndrome caused by mutations in the zinc finger homeo box 1B, *Am. J. Med. Genet.*, **108**, 177–81.
31. Lantinga-van Leeuwen, I.S., Leonhard, W.N., Dauwerse, H., et al. 2005, Common regulatory elements in the polycystic kidney disease 1 and 2 promoter regions, *Eur. J. Hum. Genet.*, **13**, 649–59.
32. Bogni, A., Cheng, C., Liu, W., et al. 2006, Genome-wide approach to identify risk factors for therapy-related myeloid leukemia, *Leukemia*, **20**, 239–46.